13

RÉGRESSION MULTIPLE

STATISTIQUES APPLIQUÉES

dunnhumby[*]
London, Royaume-Uni

Fondée en 1989 par le couple Clive Humby (mathématicien) et Edwina Dunn (expert en marketing), dunnhumby combinent des capacités naturelles à de grandes idées pour identifier et justifier les comportements d'achats de consommateurs. La société transforme ces informations en stratégies qui génèrent de la croissance et une loyauté à toute épreuve, améliorant *in fine* la valeur de marque et l'expérience client.

Employant plus de 950 personnes en Europe, en Asie et en Amérique, dunnhumby est au service de nombreuses sociétés de renom comme Kroger, Tesco, Coca-Cola, General Mill, Kimberley-Clark, PepsiCo, Procter&Gamble et Home Depot. dunnhumbyUSA et la société Kroger forment une entreprise commune qui a ses bureaux à New York, Chicago, Atlanta, Minneapolis, Cincinnati et Portland.

Les recherches effectuées par la société commencent par la collecte de données sur les clients de ses clients. Les données proviennent des cartes de fidélité, des caisses automatiques et d'études de marché traditionnelles. L'analyse des données permet de traduire des milliards de données individuelles en informations détaillées sur le comportement, les préférences et le style de vie des clients. De telles informations permettent de mettre en place des programmes de vente plus pertinents, de faire de recommandations en matière de stratégies tarifaires, de promotion et d'assortiments de produits.

Les chercheurs ont utilisé une technique de régression multiple appelée régression logistique pour analyser les données des clients. En utilisant la régression logistique, une estimation de l'équation de régression multiple de la forme suivante a été développée.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_p x_p$$

La variable dépendante $\hat{y}$ est une prévision de la probabilité qu'un client appartienne à un groupe de clients particulier. Les variables indépendantes $x_1, x_2, x_3, \ldots, x_p$ sont des mesures du comportement d'achat réel du client et peuvent inclure le type de produits achetés, le jour de la semaine, l'heure, etc. L'analyse permet d'identifier les variables indépendantes qui sont les plus pertinentes pour prédire à quel groupe appartient ce client et mieux comprendre la population de clients, ce qui permet ensuite d'effectuer des analyses plus approfondies avec une plus grande confiance. L'objectif de l'analyse est de comprendre le client dans le but de développer des offres, des politiques marketing qui maximiseront la pertinence des services proposés à chaque groupe de clients.

Dans ce chapitre, nous introduirons la régression multiple et montrerons comment les concepts de la régression linéaire simple introduits au chapitre 12 peuvent être étendus à une régression multiple. De plus, nous montrerons comment utiliser les logiciels informatiques pour effectuer des régressions multiples. Dans la dernière section du chapitre, nous introduirons la régression logistique en utilisant un exemple qui illustre comment cette technique est utilisée en marketing.

* Les auteurs remercient Paul Hunter, vice-président de Solutions pour dunnhumby de leur avoir fourni ce Statistiques appliquées.

Dans le chapitre 12, nous avons présenté l'analyse de la régression linéaire simple et illustré son application au travers d'une équation estimée de la régression qui décrit la relation entre deux variables. Pour mémoire, la variable expliquée est appelée variable dépendante et la variable explicative est appelée variable indépendante. Dans ce chapitre,

nous poursuivons notre étude de l'analyse de la régression en considérant des situations impliquant au moins deux variables indépendantes. Il s'agit de l'analyse de la régression multiple, qui nous permet de considérer plus de facteurs et donc d'obtenir de meilleures estimations que dans le cadre d'une régression linéaire simple.

13.1 LE MODÈLE DE RÉGRESSION MULTIPLE

L'analyse de la régression multiple est l'étude de la relation entre une variable dépendante y et au moins deux variables indépendantes. Dans le cas général, nous noterons p le nombre de variables indépendantes.

13.1.1 Modèle de régression et équation de la régression

Les concepts de modèle de régression et d'équation de la régression, introduits dans le chapitre précédent, sont applicables au cas multiple. L'équation qui décrit comment est reliée la variable dépendante y aux variables indépendantes $x_1, x_2, \ldots, x_p$ et à un terme d'erreur, est appelée **modèle de régression multiple**. Nous supposons pour commencer que le modèle de régression multiple est de la forme suivante.

► **Modèle de régression multiple**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \qquad (13.1)$$

Dans le modèle de régression multiple, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ sont les paramètres de la population et le terme d'erreur ε (la lettre grecque epsilon) est une variable aléatoire. Un examen approfondi de ce modèle révèle que y est une fonction linéaire de $x_1, x_2, \ldots, x_p$ (la partie $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$) plus un terme d'erreur ε. Le terme d'erreur prend en compte la variabilité de y qui n'est pas expliquée par l'impact linéaire des p variables indépendantes.

Dans la section 13.4, nous discuterons des hypothèses d'un modèle de régression multiple et du terme d'erreur ε. L'une des hypothèses est que la moyenne ou espérance mathématique de ε est nulle. Par conséquent, la moyenne ou espérance mathématique de y, notée $E(y)$, est égale à $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. L'équation qui décrit comment la moyenne de y est liée à $x_1, x_2, \ldots, x_p$ est appelée **l'équation de la régression multiple**.

► **Équation de la régression multiple**

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (13.2)$$

13.1.2 Équation estimée de la régression multiple

Si les valeurs de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ étaient connues, l'expression (13.2) pourrait être utilisée pour calculer la moyenne de y pour des valeurs données de $x_1, x_2, \ldots, x_p$. Malheureusement, ces paramètres ne sont généralement pas connus et doivent être estimés à partir des données d'un échantillon. On utilise un échantillon aléatoire simple pour calculer les statistiques

d'échantillon $b_0, b_1, b_2, \ldots, b_p$ utilisées comme estimateurs ponctuels des paramètres de la population $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Ces statistiques d'échantillon fournissent **l'équation estimée de la régression multiple** suivante.

► **Équation estimée de la régression multiple**

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p \qquad (13.3)$$

où

$b_0, b_1, b_2, \ldots, b_p$ sont les estimations de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ et $\hat{y}$ correspond à la valeur estimée de la variable dépendante.

La figure 13.1 illustre le processus d'estimation dans le cadre d'une régression multiple.

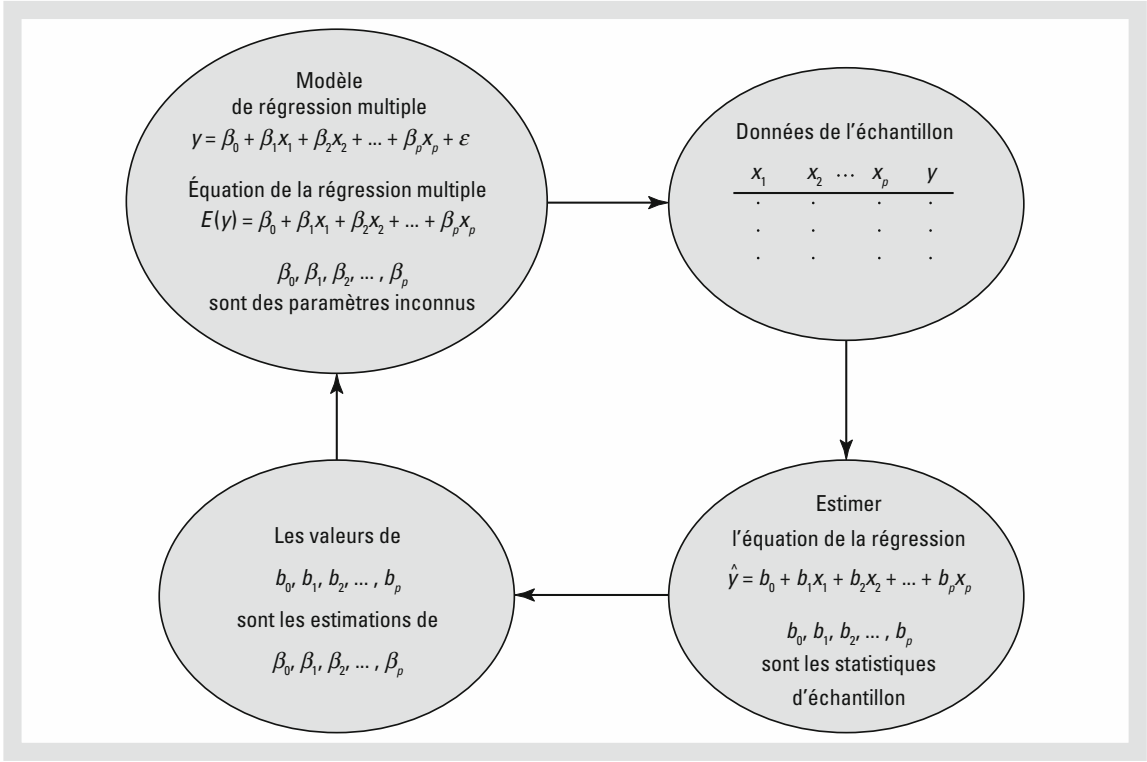

Figure 13.1 ***Processus d'estimation dans le cadre d'une régression multiple***

Dans le cadre d'une régression linéaire simple, b_0 et b_1 étaient les statistiques d'échantillon utilisées pour estimer les paramètres β_0 et β_1. L'analyse de la régression multiple est le pendant de cette inférence statistique, $b_0, b_1, b_2, \ldots, b_p$ étant les statistiques d'échantillon utilisées pour estimer les paramètres $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

13.2 LA MÉTHODE DES MOINDRES CARRÉS

Dans le chapitre 12, nous avons utilisé la **méthode des moindres carrés** pour estimer l'équation de la régression qui constitue la meilleure approximation d'une relation linéaire entre les variables dépendante et indépendante. Cette même approche est utilisée pour estimer l'équation de la régression multiple. Le critère des moindres carrés est reformulé ici.

► **Critère des moindres carrés**

$$\min \sum (y_i - \hat{y}_i)^2 \tag{13.4}$$

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

Les valeurs estimées de la variable dépendante sont calculées en utilisant l'équation estimée de la régression multiple,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Comme l'indique l'expression (13.4), la méthode des moindres carrés se sert des données de l'échantillon pour obtenir les valeurs de b_0, b_1, b_2, …, b_p qui minimisent la somme des carrés des résidus (les écarts entre les valeurs observées (y_i) et les valeurs estimées ($\hat{y}_i$) de la variable dépendante).

Dans le chapitre 12, nous avons présenté les formules de calcul des estimateurs des moindres carrés b_0 et b_1 dans le cadre de l'équation estimée de la régression linéaire simple $\hat{y} = b_0 + b_1 x$. Pour des ensembles de données relativement petits, nous étions capables d'utiliser ces formules pour calculer, à la main, b_0 et b_1. Par contre, dans le cadre d'une régression multiple, la présentation des formules de calcul des coefficients de régression b_0, b_1, b_2, …, b_p nécessite l'utilisation de l'algèbre matriciel et s'écarte de l'objet de cet ouvrage. Par conséquent, nous nous focaliserons sur l'utilisation des logiciels pour obtenir l'équation estimée de la régression multiple ainsi que d'autres informations. L'accent sera mis sur l'interprétation des résultats de la programmation plutôt que sur les calculs proprement dits de la régression.

13.2.1 Un exemple : la société de transport Butler

Pour illustrer l'analyse de la régression multiple, nous considérons un problème rencontré par la société de transport Butler, implantée en Californie du Sud. La société Butler effectue des livraisons locales. Pour améliorer les plannings de travail, les responsables souhaitent estimer la durée quotidienne des trajets effectués par les chauffeurs.

Les responsables supposaient initialement que la durée totale des trajets quotidiens était fortement liée au nombre de kilomètres parcourus pour effectuer les livraisons.

Un échantillon aléatoire simple de dix livraisons a fourni les données présentées dans le tableau 13.1 (cf. fichier en ligne Butler) et le nuage de point représenté à la figure 13.2. Au regard de ce nuage de point, les responsables ont supposé que le modèle de régression linéaire simple $y = \beta_0 + \beta_1 x_1 + \varepsilon$ pouvait être utilisé pour décrire la relation entre la durée totale des trajets (y) et le nombre de kilomètres parcourus (x_1). Pour estimer les paramètres β_0 et β_1, ils ont utilisé la méthode des moindres carrés afin d'obtenir l'équation estimée de la régression

$$\hat{y} = b_0 + b_1 x_1 \tag{13.5}$$

La figure 13.3 correspond au résultat de la programmation sous Minitab d'une régression linéaire simple, obtenu en utilisant les données du tableau 13.1. L'équation estimée de la régression est

$$\hat{y} = 1{,}27 + 0{,}0678 x_1$$

Au seuil de signification $\alpha = 0{,}05$, la valeur F égale à 15,81 et la valeur p associée à cette statistique de test, égale à 0,004, indiquent que la relation est significative ; on peut donc rejeter $H_0 : \beta_1 = 0$, la valeur p étant inférieure à α égal à 0,05. Notez qu'on obtient la même conclusion en utilisant la valeur t, égale à 3,98 et la valeur p qui lui est associée, égale à 0,004. Ainsi, nous pouvons conclure que la relation entre la durée totale des trajets et le nombre de kilomètres parcourus est significative ; des durées de trajets plus longues sont associées à un plus grand nombre de kilomètres parcourus. Puisque le coefficient de détermination (exprimé en pourcentage) est égal à 66,4 %, 66,4 % de la variabilité de la durée des trajets peut être expliquée linéairement par le nombre de kilomètres parcourus. Ce résultat est acceptable, mais les responsables souhaitent ajouter une seconde variable indépendante pour expliquer la variabilité restante de la variable dépendante.

Tableau 13.1 ***Données préliminaires de la société Butler***

Permis de conduire	x_1 = Kilomètres parcourus	y = Temps de trajet (heures)
1	100	9,3
2	50	4,8
3	100	8,9
4	100	6,5
5	50	4,2
6	80	6,2
7	75	7,4
8	65	6,0
9	90	7,6
10	90	6,1

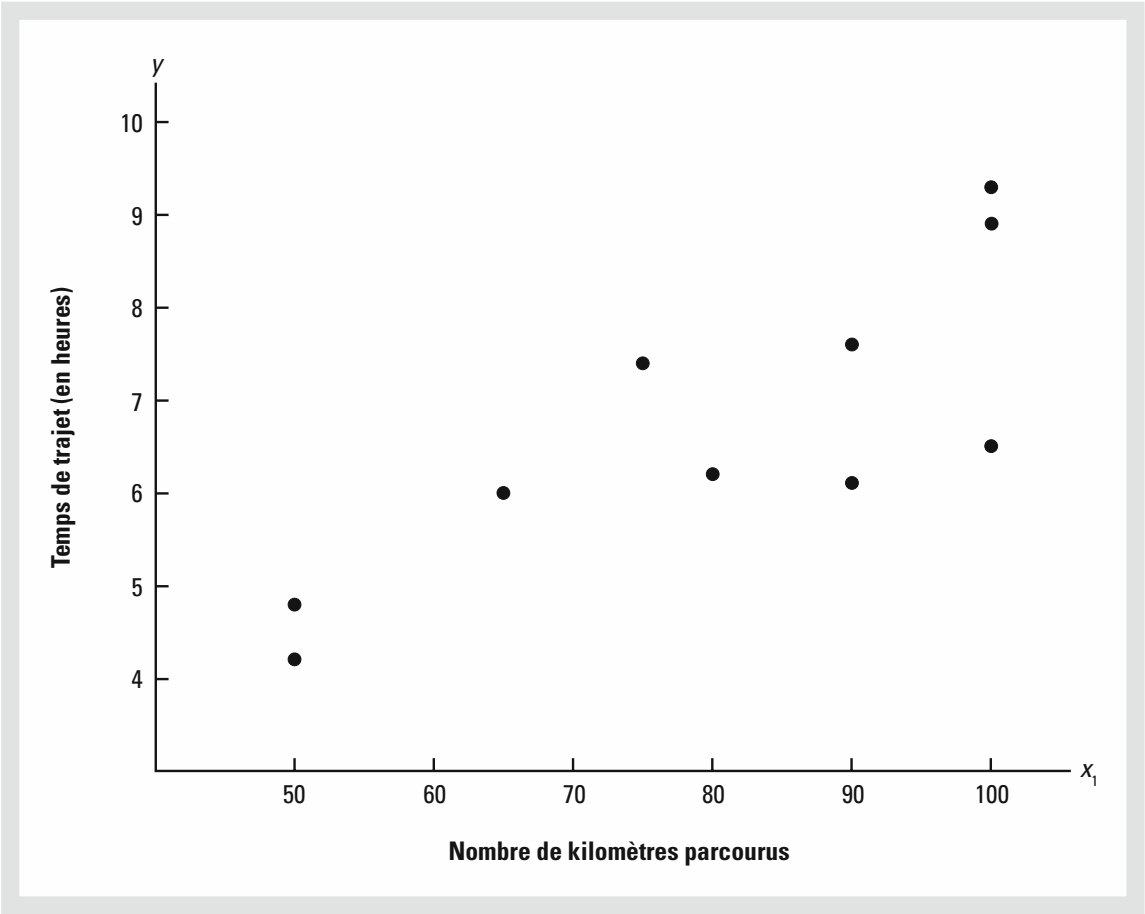

Figure 13.2 ***Nuage de points des données préliminaires de la société Butler***

En essayant d'identifier une autre variable indépendante, les responsables ont pensé que le nombre de livraisons pouvait également expliquer la durée totale du trajet. Les données de la société Butler, y compris celles sur le nombre de livraisons effectuées, sont présentées dans le tableau 13.2. Le résultat de la programmation sous Minitab, en considérant le nombre de kilomètres parcourus (x_1) et le nombre de livraisons effectuées (x_2) en tant que variables indépendantes, est reproduit à la figure 13.4. L'équation estimée de la régression est

$$\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2 \qquad (13.6)$$

Dans la section suivante, nous discuterons de l'utilisation du coefficient de détermination multiple pour mesurer l'adéquation de cette équation estimée de la régression aux données. Tout d'abord, examinons plus attentivement les valeurs de $b_1 = 0{,}0611$ et $b_2 = 0{,}923$ dans l'équation (13.6).

```
The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor       Coef   SE Coef      T      p
Constant       1.274     1.401   0.91  0.390
Miles        0.06783   0.01706   3.98  0.004

S = 1.002    R-sq = 66.4%    R-sq(adj) = 62.2%

Analysis of Variance

SOURCE           DF       SS       MS       F      p
Regression        1   15.871   15.871   15.81  0.004
Residual Error    8    8.029    1.004
Total             9   23.900
```

Figure 13.3 *Output Minitab de l'exemple de la société Butler avec une variable indépendante*

Le nom des variables apparaissant dans l'output Minitab (Miles pour kilomètres et Time pour durée des trajets) a été entré dans la feuille de calcul.

Tableau 13.2 *Données pour l'exemple Butler avec le nombre de kilomètres parcourus (x_1) et le nombre de livraisons effectuées (x_2) considérés comme variables indépendantes*

Permis de conduire	x_1 = Kilomètres parcourus	x_2 = Livraisons effectuées	y = Temps de trajet (heures)
1	100	4	9,3
2	50	3	4,8
3	100	4	8,9
4	100	2	6,5
5	50	2	4,2
6	80	2	6,2
7	75	3	7,4
8	65	4	6,0
9	90	3	7,6
10	90	2	6,1

```
The regression equation is
Time = -0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor         Coef    SE Coef       T       p
Constant       -0.8687     0.9515   -0.91   0.392
Miles         0.061135   0.009888    6.18   0.000
Deliveries      0.9234     0.2211    4.18   0.004

S = 0.5731    R-sq = 90.4%    R-sq(adj) = 87.6%

Analysis of Variance

SOURCE           DF       SS       MS       F       p
Regression        2   21.601   10.800   32.88   0.000
Residual Error    7    2.299    0.328
Total             9   23.900
```

Figure 13.4 ***Output Minitab de l'exemple de la société Butler avec deux variables indépendantes***

Le nom des variables apparaissant dans l'output Minitab (Miles pour le nombre de kilomètres parcourus, Deliveries pour le nombre de livraisons effectuées et Time pour la durée des trajets) a été entré dans la feuille de calcul.

Les étapes de programmation sous Minitab nécessaires pour générer l'output présenté à la figure 13.4 sont fournies dans l'annexe 13.1.

13.2.2 Remarque sur l'interprétation des coefficients

Une observation peut être faite sur la relation entre l'équation estimée de la régression avec une seule variable indépendante, le nombre de kilomètres parcourus, et l'équation qui comprend deux variables indépendantes, le nombre de kilomètres parcourus et le nombre de livraisons effectuées. La valeur de b_1 n'est pas identique dans les deux cas. Dans une régression linéaire simple, nous interprétons b_1 comme une estimation de l'effet sur y d'une variation d'une unité de la variable indépendante. Dans une analyse de régression multiple, cette interprétation est légèrement modifiée. Dans une analyse de régression multiple, chaque coefficient est interprété de la façon suivante : b_i représente une estimation d'un changement de y suite à un changement d'une unité de x_i lorsque toutes les autres variables indépendantes sont constantes. Dans l'exemple de la société de transport Butler impliquant deux variables indépendantes, b_1 est égal à 0,0611. Ainsi, 0,0611 heure est une estimation de l'augmentation attendue de la durée des trajets suite à une augmentation de la distance parcourue d'un kilomètre, lorsque le nombre de livraisons reste constant.

De même, puisque b_2 est égal à 0,923, 0,923 heure est une estimation de l'augmentation attendue de la durée des trajets suite à une livraison supplémentaire, lorsque le nombre de kilomètres parcourus reste constant.

EXERCICES

Remarque à l'attention des étudiants : Ces exercices ont été élaborés pour être résolus en utilisant un logiciel statistique.

Méthode

1. L'équation de la régression d'un modèle composé de deux variables indépendantes estimée à partir de dix observations s'écrit :

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

a) Interpréter b_1 et b_2 dans cette équation estimée de la régression.

b) Estimer y lorsque $x_1 = 180$ et $x_2 = 310$.

2. Considérez les données suivantes (cf. fichier en ligne Exo2) relatives à une variable dépendante y et deux variables indépendantes, x_1 et x_2.

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

a) Utiliser ces données pour estimer l'équation de la régression reliant y à x_1. Estimer y si $x_1 = 45$.

b) Utiliser ces données pour estimer l'équation de la régression reliant y à x_2. Estimer y si $x_2 = 15$.

c) Utiliser ces données pour estimer l'équation de la régression reliant y à x_1 et x_2. Estimer y si $x_1 = 45$ et $x_2 = 15$.

3. Dans une analyse de la régression faite à partir de 30 observations, on a estimé l'équation de la régression suivante.

$$\hat{y} = 17{,}6 + 3{,}8x_1 - 2{,}3x_2 + 7{,}6x_3 + 2{,}7x_4$$

a) Interpréter b_1, b_2, b_3 et b_4 dans cette équation estimée de la régression.

b) Estimer y lorsque $x_1 = 10$, $x_2 = 5$, $x_3 = 1$ et $x_4 = 2$.

Applications

4. Un magasin de chaussures a estimé l'équation de la régression suivante reliant les ventes au stock de marchandises et aux dépenses publicitaires.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

où x_1 correspond au stock (en milliers de dollars), x_2 aux dépenses publicitaires (en milliers de dollars) et y aux ventes (en milliers de dollars).

a) Estimer les ventes résultant d'un stock de 15 000 dollars et d'un budget publicitaire de 10 000 dollars.

b) Interpréter b_1 et b_2 dans cette équation estimée de la régression.

5. Le propriétaire de la société Showtime Movie Theaters voudrait estimer le chiffre d'affaires hebdomadaire en fonction des dépenses publicitaires. Les données historiques d'un échantillon de huit semaines sont présentées dans le tableau ci-dessous (cf. fichier en ligne Showtime).

Chiffre d'affaires hebdomadaire (milliers de dollars)	Publicité télévisée (milliers de dollars)	Publicité dans les journaux (milliers de dollars)
96	5,0	1,5
90	2,0	2,0
95	4,0	1,5
92	2,5	2,5
95	3,0	3,3
94	3,5	2,3
94	2,5	4,2
94	3,0	2,5

a) Estimer l'équation de la régression en considérant le montant des dépenses publicitaires télévisées comme variable indépendante.

b) Estimer l'équation de la régression en considérant les dépenses publicitaires télévisées et dans les journaux comme variables indépendantes.

c) Est-ce que le coefficient de l'équation estimée de la régression associé aux dépenses publicitaires télévisées est le même dans les questions (a) et (b) ? Interpréter le coefficient dans chaque cas.

d) Quelle est l'estimation du revenu brut d'une semaine lorsque 3 500 dollars sont dépensés en publicité télévisée et 1 800 dollars en publicité dans les journaux.

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 l'association (Association), le nombre moyen de yards parcourus en faisant des passes

(yards), le nombre de lancers interceptés (Interceptions) et le pourcentage de parties gagnées (% parties gagnées) ont été collectées à partir d'un échantillon aléatoire de 16 équipes de la NFL au cours de la saison 2011 (site Internet de la NFL, 12 février 2012).

Équipe	Association	Yards	Interceptions	% parties gagnées
Arizona Cardinals	NFC	6,5	0,042	50,0
Atlanta Falcons	NFC	7,1	0,022	62,5
Carolina Panthers	NFC	7,4	0,033	37,5
Cincinnati Bengals	AFC	6,2	0,026	56,3
Detroit Lions	NFC	7,2	0,024	62,5
Green Bay Packers	NFC	8,9	0,014	93,8
Houstan Texans	AFC	7,5	0,019	62,5
Indianapolis Colts	AFC	5,6	0,026	12,5
Jacksonville Jaguars	AFC	4,6	0,032	31,3
Minnesota Vikings	NFC	5,8	0,033	18,8
New England Patriots	AFC	8,3	0,020	81,3
New Orleans Saints	NFC	8,1	0,021	81,3
Oakland Raiders	AFC	7,6	0,044	50,0
San Francisco 49ers	NFC	6,5	0,011	81,3
Tennessee Titans	AFC	6,7	0,024	56,3
Washington Redskins	NFC	6,4	0,041	31,3

a) Développer une équation estimée de la régression qui permettrait de prévoir le pourcentage de parties gagnées étant donné le nombre moyen de yards parcourus en faisant des passes.

b) Développer une équation estimée de la régression qui permettrait de prévoir le pourcentage de parties gagnées étant donné le nombre de lancers interceptés.

c) Développer une équation estimée de la régression qui permettrait de prévoir le pourcentage de parties gagnées étant donnés le nombre moyen de yards parcourus en faisant des passes et le nombre de lancers interceptés.

d) Le nombre moyen de yards parcourus en faisant des passes par les Kansas City Chiefs fut de 6,2 et le nombre de lancers interceptés de 0,036. Utiliser l'équation de la régression estimée obtenue à la question (c) 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. *PC World* a évalué quatre caractéristiques de 10 ordinateurs ultra-portables : les caractéristiques techniques, la performance, le design et le prix. Chaque caractéristique était évaluée sur une échelle allant de 1 à 100 points. Une note globale a ensuite été attribuée à chaque ordinateur. Le tableau suivant (cf. fichier en ligne Ordinateur portable) fournit l'évaluation de la performance, l'évaluation des caractéristiques techniques et la note globale des 10 ordinateurs ultra-portables (site internet de *PC World*, 5 février 2009).

Modèle	Évaluation de la performance	Évaluation des caractéristiques techniques	Note globale
Thinkpad X200	77	87	83
VGN-Z598U	97	85	82
U6V	83	80	81
Elitebook 2530P	77	75	78
X360	64	80	78
Thinkpad X300	56	76	78
Ideapad U110	55	81	77
Micro Express JFT2500	76	73	75
Thoughbook W7	46	79	73
HP Voodoo Envy 133	54	68	72

a) Développer l'équation estimée de la régression permettant de prévoir la note globale en fonction de l'évaluation de la performance.

b) Développer l'équation estimée de la régression permettant de prévoir la note globale en fonction de l'évaluation de la performance et de l'évaluation des caractéristiques techniques.

c) Prévoir la note globale d'un ordinateur dont la performance s'élève à 80 et les caractéristiques techniques à 70.

8. La liste Or 2012 de *Condé Nast Traveler* a fourni les évaluations des 20 meilleures croisières en bateau (site Internet de *Condé Nast Traveler*, 1er mars 2012). Les données reprises ci-dessous (cf. fichier en ligne Bateau) correspondent aux notes attribuées à chaque bateau de croisière, fondées sur les résultats de l'enquête annuelle Readers' Choice menée par *Condé Nast Traveler*. Chaque note représente le pourcentage de personnes interrogées qui ont évalué le bateau comme excellent ou très bon selon plusieurs critères comme les excursions sur le littoral et les repas. Une note globale est également reportée et utilisée pour classer les bateaux. Le premier bateau du classement, le Seabourn Odyssey, a obtenu une note globale de 94,4, et la note associée aux repas la plus élevée à 97,8.

Bateaux	Note globale	Excursions sur le littoral	Repas
Seabourn Odyssey	94,4	90,9	97,8
Seabourn Pride	93,0	84,2	96,7
National Geographie Endeavor	92,9	100,0	88,5
Seabourn Sojourn	91,3	94,8	97,1
Paul Gauguin	90,5	87,9	81,2
Seabourn Legend	90,3	82,1	98,8
Seabourn Spirit	90,2	86,3	92,0
Silver Explorer	89,9	92,6	88,9
Silver Spirit	89,4	85,9	90,8
Seven Seas Navigator	89,2	83,3	90,5
Silver Whisperer	89,2	82,0	88,6

Bateaux	Note globale	Excursions sur le littoral	Repas
National Geographic Explorer	89,1	93,1	89,7
Silver Cloud	88,7	78,3	91,3
Celebrity Xpedition	87,2	91,7	73,6
Silver Shadow	87,2	75,0	89,7
Silver Wind	86,6	78,1	91,6
SeaDream II	86,2	77,4	90,9
Wind Star	86,1	76,5	91,5
Wind Surf	86,1	72,3	89,3
Wind Spirit	85,2	77,4	91,9

a) Développer l'équation estimée de la régression qui permettrait de prévoir la note globale étant donnée la note attribuée aux excursions.

b) Considérer l'ajout de la variable indépendante relative aux repas. Développer l'équation estimée de la régression qui permettrait de prévoir la note globale étant données les notes attribuées aux excursions et aux repas.

c) Estimer la note globale d'un bateau de croisière dont les excursions sont notées 80 et les repas 90.

9. L'Association des golfeurs professionnels (PGA) conserve des données sur les performances et les gains des participants au tournoi PGA. Au cours de la saison 2012, Bubba Watson a supplanté tous les joueurs en termes de distance de frappe, avec une moyenne de 309,2 yards par frappe. Les facteurs influençant la distance de frappe sont la vitesse à laquelle le club touche la balle, la vitesse de la balle envoyée et l'angle de frappe (l'angle vertical de la balle immédiatement après avoir été touchée par le club). Au cours de la saison 2012, la vitesse moyenne du club de Bubba Watson fut de 124,69 miles par heure, la vitesse moyenne de ses balles de 184,98 miles par heure et un angle moyen de frappe de 8,79 degrés. Le fichier en ligne intitulé PGADrivingDist contient les données sur les distances de frappe et ces différents facteurs pour 190 participants au tournoi PGA (site Internet du PGA Tour, 1er novembre 2012).

a) Développer l'équation estimée de la régression qui pourrait être utilisée pour prévoir le nombre moyen de yards parcourus par la balle étant donnée la vitesse à laquelle le club a touché la balle.

b) Développer l'équation estimée de la régression qui pourrait être utilisée pour prévoir le nombre moyen de yards parcourus par la balle étant donnée la vitesse de la balle envoyée.

c) Il a été recommandé d'utiliser à la fois la vitesse à laquelle le club a touché la balle et la vitesse de la balle envoyée pour prévoir le nombre moyen de yards parcourus par la balle. Êtes-vous d'accord ? Expliquer.

d) Développer l'équation estimée de la régression qui pourrait être utilisée pour prévoir le nombre moyen de yards parcourus par la balle étant donnée la vitesse de la balle envoyée et l'angle de frappe.

e) Supposez qu'un nouveau participant au tournoi de 2013 ait une vitesse de balle de 170 miles par heure et un angle de frappe de 11 degrés. Utiliser l'équation estimée

de la régression obtenue à la question (d) pour prévoir le nombre moyen de yards parcourus par la balle frappée par ce joueur.

10. La ligue principale de baseball (MLB) est constituée des équipes qui participent à la Ligue américaine et à la Ligue nationale. La MLB collecte diverses statistiques sur les équipes et les joueurs. Certaines des statistiques souvent utilisées pour évaluer la qualité des lancers sont les suivantes :

Buts : Le nombre de buts sur balles par 9 manches lancées
SO/Manche : Le nombre moyen de strikeouts par manche lancée
HR/Manche : Le nombre moyen de home runs par manche lancée
Coups sûrs/Manche : Le nombre de coups sûrs par manche lancée

Les données suivantes (cf. fichier en ligne MLB) indiquent les valeurs de ces statistiques pour un échantillon aléatoire de 20 lanceurs appartenant la ligue américaine durant la saison 2011 (site Internet de la MLB, 1er mars 2012).

Joueur	Équipe	W	L	Buts	SO/Manche	HR/Manche	Coups sûrs/Manche
Verlander, J	DET	24	5	2,40	1,00	0,10	0,29
Beckett, J	BOS	13	7	2,89	0,91	0,11	0,34
Wilson, C	TEX	16	7	2,94	0,92	0,07	0,40
Sabathia, C	NYY	19	8	3,00	0,97	0,07	0,37
Haren, D	LAA	16	10	3,17	0,81	0,08	0,38
McCarthy, B	OAK	9	9	3,32	0,72	0,06	0,43
Santana, E	LAA	11	12	3,38	0,78	0,11	0,42
Lester, J	BOS	15	9	3,47	0,95	0,10	0,40
Hernandez, F	SEA	14	14	3,47	0,95	0,08	0,42
Buehrle, M	CWS	13	9	3,59	0,53	0,10	0,45
Pineda, M	SEA	9	10	3,74	1,01	0,11	0,44
Colon, B	NYY	8	10	4,00	0,82	0,13	0,52
Tomlin, J	CLE	12	7	4,25	0,54	0,15	0,48
Pavano, C	MIN	9	13	4,30	0,46	0,10	0,55
Danks, J	CWS	8	12	4,33	0,79	0,11	0,52
Guthrie, J	BAL	9	17	4,33	0,63	0,13	0,54
Lewis, C	TEX	14	10	4,40	0,84	0,17	0,51
Scherzer, M	DET	15	9	4,43	0,89	0,15	0,52
Davis, W	TB	11	10	4,45	0,57	0,13	0,52
Porcello, R	DET	14	9	4,75	0,57	0,10	0,57

a) Développer l'équation estimée de la régression qui peut être utilisée pour prévoir le nombre moyen de coups sûrs par manche étant donné le nombre moyen de strikeouts par manche.

b) Développer l'équation estimée de la régression qui peut être utilisée pour prévoir le nombre moyen de coups sûrs par manche étant donné le nombre moyen de home runs par manche.

c) Développer l'équation estimée de la régression qui peut être utilisée pour prévoir le nombre moyen de coups sûrs par manche étant donnés les nombres moyens de strikeouts et de home runs par manche.

d) A.J. Burnett, un lanceur des New York Yankees, a à son actif un nombre moyen de strikeouts par manche de 0,91 et un nombre moyen de home runs par manche de 0,16. Utiliser l'équation estimée de la régression obtenue à la question (c) pour prévoir le nombre moyen de coups sûrs par manche de A.J. Burnett (remarque : la vraie valeur est de 0,6).

e) Il a été suggéré d'utiliser également le nombre moyen de buts comme autre variable indépendante à la question (c). Que pensez-vous de cette suggestion ?

13.3 LE COEFFICIENT DE DÉTERMINATION MULTIPLE

Dans le cadre d'une régression linéaire simple, nous avons montré que la somme totale des carrés pouvait être divisée en deux composantes : la somme des carrés de la régression et la somme des carrés des résidus. La même procédure s'applique à la somme des carrés dans le cadre d'une régression multiple.

► **Relation entre SCT, SCreg et SCres**

$$SCT = SCreg + SCres \tag{13.7}$$

où

$SCT = \sum (y_i - \bar{y})^2$ correspond à la somme des carrés totale

$SCreg = \sum (\hat{y}_i - \bar{y})^2$ correspond à la somme des carrés de la régression

$SCres = \sum (y_i - \hat{y}_i)^2$ correspond à la somme des carrés des résidus

À cause de la complexité des calculs de ces trois sommes des carrés, nous nous reposons sur les logiciels informatiques pour déterminer ces valeurs. L'analyse de la variance faite par Minitab, présentée à la figure 13.4, fournit les trois valeurs dans le cadre du problème de la société de transport Butler à deux variables indépendantes : *SCT* = 23,900, *SCreg* = 21,601 et *SCres* = 2,299. Avec une seule variable indépendante (le nombre de kilomètres parcourus), l'output de Minitab présenté à la figure 13.3 indiquait les valeurs suivantes : *SCT* = 23,900, *SCreg* = 15,871 et *SCres* = 8,029. La valeur de *SCT* est identique dans les deux cas, puisqu'elle ne dépend pas de $\hat{y}$, mais l'introduction d'une seconde variable indépendante (le nombre de livraisons) accroît *SCreg* et réduit *SCres*. En conséquence, l'équation estimée de la régression multiple est plus adaptée aux données observées.

Dans le chapitre 12, nous avons mesuré l'adéquation de l'équation estimée de la régression aux données grâce au coefficient de détermination $r^2 = SCreg / SCT$. Le même concept s'applique à la régression multiple. Le terme **coefficient de détermination multiple** indique que nous mesurons l'adéquation d'une équation estimée de régression multiple. Le coefficient de détermination multiple, noté R^2, est calculé de la façon suivante :

► **Coefficient de détermination multiple**

$$R^2 = SCreg/SCT \tag{13.8}$$

Le coefficient de détermination multiple peut être interprété comme la proportion de la variabilité de la variable dépendante expliquée par l'équation estimée de la régression multiple. En le multipliant par 100, on peut l'interpréter comme le pourcentage de la variation de y expliquée par l'équation estimée de la régression.

Dans l'exemple de la société de transport Butler à deux variables indépendantes,

$$R^2 = \frac{21{,}601}{23{,}900} = 0{,}904$$

Ainsi, 90,4 % de la variabilité du temps de trajet y est expliquée par l'équation estimée de la régression multiple, ayant pour variables indépendantes le nombre de kilomètres parcourus et le nombre de livraisons effectuées. L'output Minitab de la figure 13.4 fournit également le coefficient de détermination multiple ; il est noté $R - sq = 90{,}4$ %.

La figure 13.3 indique que la valeur du coefficient de détermination de l'équation estimée de la régression avec une seule variable indépendante, le nombre de kilomètres parcourus (x_1), est égale à 66,4 %. Ainsi, le pourcentage de la variabilité de la durée des trajets expliquée par l'équation estimée de la régression est passé de 66,4 % à 90,4 % en ajoutant le nombre de livraisons effectuées comme seconde variable indépendante. En général, R^2 augmente lorsque des variables indépendantes sont ajoutées au modèle.

Ajouter des variables indépendantes réduit l'erreur de prévision, et par conséquent, la somme des carrés des résidus. Puisque *SCreg = SCT – SCres*, lorsque SCres diminue, SCreg augmente, entraînant une augmentation de R^2 = *SCreg/SCT*.

Beaucoup d'analystes préfèrent ajuster R^2 au nombre de variables indépendantes pour éviter de surestimer l'impact de l'ajout d'une variable indépendante sur la part de la variabilité expliquée par l'équation estimée de la régression. Avec n le nombre d'observations et p le nombre de variables indépendantes, le coefficient de détermination multiple ajusté est calculé de la façon suivante :

► **Coefficient de détermination multiple ajusté**

$$R_a^2 = 1 - \left(1 - R^2\right)\frac{n - 1}{n - p - 1} \qquad (13.9)$$

Si une variable est ajoutée dans le modèle, R^2 augmente même si cette variable n'est pas statistiquement significative. Le coefficient de détermination multiple ajusté tient compte du nombre de variables indépendantes présentes dans le modèle.

Dans l'exemple de la société de transport Butler, avec $n = 10$ et $p = 2$, nous avons

$$R_a^2 = 1 - (1 - 0{,}904)\frac{10 - 1}{10 - 2 - 1} = 0{,}88$$

Ainsi, en tenant compte de la présence de deux variables indépendantes, le coefficient de détermination multiple ajusté est égal à 0,88. Cette valeur correspond à la valeur $R - sq(adj) = 87{,}6$ % dans l'output Minitab présenté à la figure 13.4. L'écart entre ces deux valeurs tient au fait que nous avons arrondi la valeur de R^2 dans notre propre calcul.

REMARQUES

Si la valeur de R^2 est faible et que le modèle contient un nombre de variables indépendantes important, le coefficient de détermination ajusté peut prendre une valeur négative. Dans de tels cas, Minitab égalise le coefficient de détermination ajusté à zéro.

EXERCICES

Méthode

11. Dans l'exercice 1, l'équation estimée de la régression suivante, fondée sur dix observations, était présentée.

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

Les valeurs de *SCT* et *SCreg* sont respectivement égales à 6 724,125 et 6 216,375.

a) Trouver *SCres*.

b) Calculer R^2.

c) Calculer R^2_a.

d) Commenter l'adéquation de la régression aux données.

12. Dans l'exercice 2, dix observations relatives à une variable dépendante y et deux variables indépendantes x_1 et x_2 étaient données. Pour celles-ci, $SCT = 15\,182{,}9$ et $SCreg = 14\,052{,}2$.

a) Calculer R^2.

b) Calculer R^2_a.

c) L'équation estimée de la régression explique-t-elle une part importante de la variabilité des données ? Expliquer.

13. Dans l'exercice 3, l'équation estimée de la régression suivante, fondée sur 30 observations, était présentée.

$$\hat{y} = 17{,}6 + 3{,}8x - 2{,}3x_2 + 7{,}6x_3 + 2{,}7x_4$$

Les valeurs de *SCT* et *SCreg* sont respectivement égales à 1 805 et 1 760.

a) Calculer R^2.

b) Calculer R^2_a.

c) Commenter l'adéquation de la régression.

Applications

14. Dans l'exercice 4, l'équation estimée de la régression suivante, reliant les ventes au stock de marchandises et aux dépenses publicitaires, était donnée.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Les données utilisées pour développer ce modèle sont issues d'une enquête auprès de dix magasins. Pour ces données $SCT = 16\,000$ et $SCreg = 12\,000$.

a) Calculer R^2.

b) Calculer R_a^2.

c) L'équation estimée de la régression explique-t-elle une part importante de la variabilité des données ? Expliquer.

15. Dans l'exercice 5 (cf. fichier en ligne Showtime), le propriétaire de la société Showtime Movie Theaters utilisait l'analyse de la régression multiple pour prévoir le chiffre d'affaires (y) en fonction des dépenses publicitaires télévisées (x_1) et dans les journaux (x_2). L'équation estimée de la régression était

$$\hat{y} = 83{,}2 + 2{,}29x_1 + 1{,}30x_2$$

Les logiciels informatiques fournissent les informations suivantes : $SCT = 25{,}5$ et $SCreg = 23{,}435$.

a) Calculer et interpréter R^2 et R_a^2.

b) Lorsque seules les dépenses publicitaires télévisées sont considérées en tant que variable indépendante, $R^2 = 0{,}653$ et $R_a^2 = 0{,}595$. Les résultats de la régression multiple sont-ils préférables ? Expliquer.

16. Dans l'exercice 6, des données (cf. fichier en ligne NFL Passes) sur le nombre moyen de yards parcourus en faisant des passes (yards), le nombre de lancers interceptés (Interceptions) et le pourcentage de parties gagnées (% parties gagnées) ont été collectées à partir d'un échantillon aléatoire de 16 équipes de la NFL au cours de la saison 2011 (site Internet de la NFL, 12 février 2012).

a) L'équation estimée de la régression qui n'utilise que le nombre moyen de yards parcourus en faisant des passes comme variable indépendante pour prévoir le pourcentage de parties gagnées, est-elle bien adaptée aux données ?

b) Discuter des bénéfices liés à l'ajout du nombre de lancers interceptés en tant que variable indépendante supplémentaire pour prévoir le pourcentage de parties gagnées.

17. Dans l'exercice 9, les données contenues dans le fichier en ligne PGADrivingDist (site Internet de PGA Tour, 1er novembre 2012) ont été utilisées pour estimer l'équation de la régression permettant de prévoir le nombre de yards parcourus par la balle (y) étant donnés la vitesse de la balle envoyée (x_1) et l'angle de frappe (x_2). L'équation estimée de la régression était $\hat{y} = 81{,}6 + 1{,}09x_1 + 1{,}65x_2$.

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

b) À la question (b) de l'exercice 9, une équation estimée de la régression a été

développée en utilisant uniquement la vitesse de la balle pour prévoir le nombre moyen de yards parcourus par la balle. L'équation estimée de la régression était $\hat{y} = 117 + 0{,}988x_1$. Comparer l'adéquation de la régression aux données obtenue en utilisant uniquement la vitesse de la balle à celle obtenue en utilisant la vitesse de la balle et l'angle de frappe.

18. Référez-vous à l'exercice 10, dans lequel les statistiques sur les lancers de la ligue principale de baseball (MLB) étaient rapportées (cf. fichier en ligne MLB) pour un échantillon aléatoire de 20 lanceurs de la ligue américaine au cours de la saison 2011 (site Internet de la MLB, 1er mars 2012).

a) À la question (c) de l'exercice 10, une équation estimée de la régression a été développée reliant le nombre moyen de coups sûrs par manche aux nombres moyens de strikeouts et de home runs par manche. Quelles sont les valeurs de R^2 et R_a^2 ?

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

c) Supposez que le nombre moyen de buts sur balles par 9 manches lancées soit utilisé comme variable dépendante à la question (c) à la place du nombre moyen de coups sûrs par manche. Est-ce que l'équation estimée de la régression qui utilise le nombre moyen de buts sur balles est mieux adaptée aux données ? Expliquer.

13.4 LES HYPOTHÈSES DU MODÈLE

Dans la section 13.1, nous avons introduit le modèle de régression multiple suivant.

► **Modèle de régression multiple**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_p x_p + \varepsilon \qquad (13.10)$$

Les hypothèses relatives au terme d'erreur ε sont le pendant de celles développées dans le cadre d'un modèle de régression linéaire simple.

► **Hypothèses sur le terme d'erreur ε dans le cadre d'un modèle de régression multiple $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_p x_p + \varepsilon$**

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

Conséquences : Pour des valeurs données de $x_1, x_2, \ldots x_p$, l'espérance mathématique de y est égale à

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_p x_p \qquad (13.11)$$

L'expression (13.11) correspond à l'équation de la régression multiple introduite dans la section 13.1. Dans cette équation, $E(y)$ représente la moyenne de toutes les valeurs possibles de y étant données les valeurs de $x_1, x_2, \ldots x_p$.

2. La variance de ε, notée σ^2 est la même pour toutes les valeurs des variables indépendantes $x_1, x_2, \ldots x_p$.

Conséquences : La variance de y le long de la droite de régression est égale à σ^2 et est la même pour toutes les valeurs de $x_1, x_2, \ldots x_p$.

3. Les valeurs de ε sont indépendantes.
Conséquences : La valeur de ε associée à une valeur particulière des variables indépendantes n'est pas liée à la valeur de ε associée à d'autres valeurs des variables indépendantes.

4. Le terme d'erreur ε est une variable aléatoire normalement distribuée, reflétant l'écart entre la valeur y et la valeur estimée de y par $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_p x_p$.
Conséquences : Puisque β_0, β_1, β_2, β_p sont constants pour des valeurs données de x_1, x_2, $\ldots$ x_p, la variable dépendante y est également une variable aléatoire normalement distribuée.

Pour approfondir l'étude de la forme de la relation exprimée par l'équation (13.11), considérez l'équation de la régression multiple à deux variables indépendantes suivante.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Le graphique de cette équation est représenté par un plan dans un espace à trois dimensions. La figure 13.5 en est une illustration. Notez que la valeur de ε indiquée correspond à la différence entre la valeur réelle de y et la valeur estimée $E(y)$ lorsque $x_1 = x_1^*$ et $x_2 = x_2^*$.

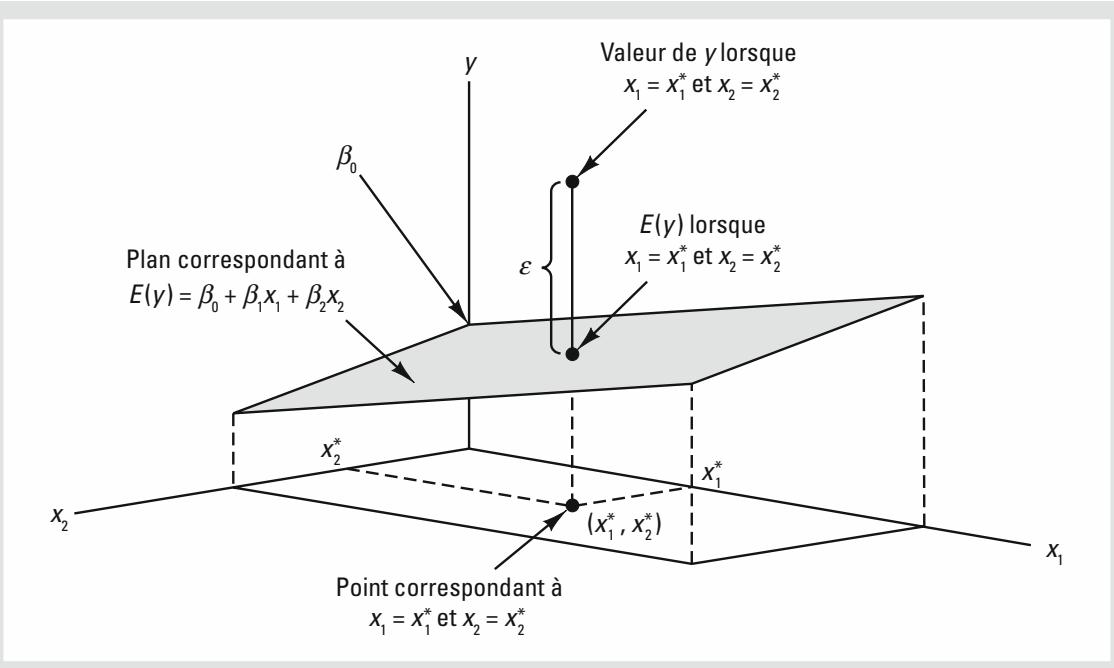

Figure 13.5 ***Graphique de l'équation de la régression dans le cadre de l'analyse d'une régression multiple à deux variables indépendantes***

Dans l'analyse de la régression, le terme *variable de réponse* est souvent utilisé à la place du terme *variable dépendante*. De plus, puisque l'équation de la régression multiple génère une surface, son graphique est appelé *surface de réponse*.

13.5 LES TESTS DE SIGNIFICATION

Dans cette section, nous montrons comment effectuer des tests de signification dans le cadre d'une relation de régression multiple. Les tests de signification utilisés dans une régression linéaire simple étaient les tests *t* de Student et *F* de Fisher. Dans le cadre d'une régression linéaire simple, les deux tests aboutissent à la même conclusion ; c'est-à-dire, si l'hypothèse nulle est rejetée, nous concluons que $\beta_1 \neq 0$. Dans le cadre d'une régression multiple, les tests de Student et de Fisher n'ont pas le même objectif.

1. Le test *F* de Fisher est utilisé pour déterminer s'il existe une relation significative entre la variable dépendante et l'ensemble des variables indépendantes ; on parle de *test de signification globale*.
2. Le test *t* de Student est utilisé pour déterminer si chacune des variables indépendantes est significative. Un test de Student est effectué pour chaque variable indépendante du modèle ; on parle de *test de signification individuelle*.

Dans la suite, nous explicitons les tests de Student et de Fisher et appliquons chacun d'entre eux au problème de régression multiple de la société de transport Butler.

13.5.1 Test de Fisher

Le modèle de régression multiple tel que défini dans la section 13.4 est

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Les hypothèses du test de Fisher concernent les paramètres du modèle de régression multiple.

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a : Au moins un des paramètres n'est pas égal à zéro

Si H_0 est rejetée, le test nous permet de conclure qu'au moins un des paramètres n'est pas égal à zéro et que la relation globale entre *y* et l'ensemble des variables indépendantes $x_1, x_2, \ldots, x_p$ est significative. Cependant, si H_0 ne peut être rejetée, nous ne disposons pas de preuves statistiques suffisantes pour conclure à l'existence d'une relation significative.

Avant de décrire les étapes d'un test de Fisher, nous devons revoir le concept de *moyenne des carrés*. La moyenne des carrés est une somme de carrés divisée par le nombre de degrés de liberté correspondant. Dans le cas d'une régression multiple, la somme des carrés totale (*SCT*) a $n - 1$ degrés de liberté, la somme des carrés de la régression (*SCreg*) a *p* degrés de liberté et la somme des carrés des résidus (*SCres*) a $n - p - 1$ degrés de liberté.

Par conséquent, la moyenne des carrés de la régression (*MCreg*) et la moyenne des carrés des résidus (*MCres*) sont respectivement égales à

$$MCreg = \frac{SCreg}{p} \tag{13.12}$$

et

$$MCres = \frac{SCres}{n - p - 1} \tag{13.13}$$

Comme nous l'avons vu au chapitre 12, *MCres* constitue un estimateur sans biais de σ^2, la variance du terme d'erreur ε. Si $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ est vraie, *MCreg* constitue également un estimateur sans biais de σ^2, et la valeur de *MCreg* / *MCres* est proche de 1. Cependant, si H_0 est fausse, *MCreg* surestime σ^2 et la valeur de *MCreg* / *MCres* augmente. Pour déterminer à partir de quelle valeur de *MCreg* / *MCres* l'hypothèse nulle peut être rejetée, nous nous basons sur le fait que si H_0 est vraie et si les hypothèses sur le modèle de régression multiple sont validées, la distribution d'échantillonnage de *MCreg* / *MCres* suit une loi de Fisher avec p degrés de liberté au numérateur et $n - p - 1$ degrés de liberté au dénominateur. Un résumé du test de signification de Fisher dans le cadre d'une régression multiple suit.

► **Test de signification globale de Fisher**

$H_0 : \beta_1 = \beta_2 \ldots = \beta_p = 0$
H_a : Au moins un des paramètres n'est pas égal à zéro

► **Statistique de test**

$$F = \frac{MCreg}{MCres} \tag{13.14}$$

► **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 loi de Fisher à p degrés de liberté au numérateur et $n - p - 1$ degrés de liberté au dénominateur.

Appliquons le test de Fisher au cas de la société de transport Butler. Avec deux variables indépendantes, les hypothèses sont écrites de la façon suivante :

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : \beta_1 \text{et/ou } \beta_2 \text{ n'est pas égal à zéro}$$

La figure 13.6 correspond à l'output de la régression multiple effectuée par Minitab, avec pour variables indépendantes, le nombre de kilomètres parcourus (x_1) et le nombre de livraisons effectuées (x_2). Dans la partie consacrée à l'analyse de la variance,

```
The regression equation is
Time = -0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor         Coef    SE Coef       T       p
Constant       -0.8687     0.9515   -0.91   0.392
Miles         0.061135   0.009888    6.18   0.000
Deliveries      0.9234     0.2211    4.18   0.004

S = 0.5731    R-sq = 90.4%    R-sq(adj) = 87.6%

Analysis of Variance

SOURCE            DF       SS       MS       F       p
Regression         2   21.601   10.800   32.88   0.000
Residual Error     7    2.299    0.328
Total              9   23.900
```

Figure 13.6 *Output Minitab obtenu dans le cadre de l'exemple de la société Butler avec deux variables indépendantes, le nombre de kilomètres parcourus (x_1) et le nombre de livraisons effectuées (x_2)*

on constate que *MCreg* est égale à 10,8, *MCres* est égale à 0,328. D'après l'équation (13.14), la statistique de test *F* est égale à

$$F = \frac{10,8}{0,328} = 32,9$$

Notez que la valeur *F* fournie par Minitab est égale à 32,88. La valeur diffère légèrement de la nôtre dans la mesure où nous avons arrondi les valeurs de *MCreg* et *MCres* dans nos calculs. Au seuil de signification $\alpha = 0,01$, la valeur $p = 0,000$ dans la dernière colonne du tableau d'analyse de la variance (cf. figure 13.6) indique que nous pouvons rejeter $H_0 : \beta_1 = \beta_2 = 0$ puisque la valeur *p* est inférieure à $\alpha = 0,01$. De même, la table 4 de l'annexe B révèle qu'avec deux degrés de liberté au numérateur et sept degrés de liberté au dénominateur, $F_{0,01} = 9,55$. Puisque $32,9 > 9,55$, nous rejetons $H_0 : \beta_1 = \beta_2 = 0$ et concluons qu'une relation significative existe entre la durée des trajets *y* et les deux variables indépendantes, le nombre de kilomètres parcourus et le nombre de livraisons effectuées.

Comme noté précédemment, la moyenne des carrés des résidus constitue un estimateur sans biais de σ^2, la variance du terme d'erreur ε. D'après la figure 13.6, l'estimation de σ^2 est *MCres* = 0,328. La racine carrée de *MCres* correspond à l'estimation de l'écart type du terme d'erreur. Comme défini dans la section 12.5, cet écart type est appelé erreur type de l'estimation et est noté *s*. Par conséquent, $s = \sqrt{MCres} = \sqrt{0,328} = 0,573$. Notez que la valeur de l'erreur type de l'estimation apparaît dans l'output Minitab (cf. figure 13.6).

Tableau 13.3 *Tableau ANOVA dans le cadre d'un modèle de régression multiple à p variables indépendantes*

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

Le tableau 13.3 correspond au tableau d'analyse de la variance (ANOVA) qui fournit les résultats du test de Fisher dans le cadre d'un modèle de régression multiple. La valeur de la statistique de test F apparaît dans la dernière colonne et peut être comparée à F_α avec p degrés de liberté au numérateur et $n - p - 1$ degrés de liberté au dénominateur, afin d'obtenir la conclusion du test d'hypothèses. En revenant à la figure 13.6, représentant l'output Minitab dans le cadre du problème de la société de transport Butler, on constate que le tableau d'analyse de la variance de Minitab contient cette information. De plus, Minitab fournit la valeur p associée à la statistique de test F.

13.5.2 Test de Student

Si le test de Fisher prouve que la relation de régression multiple est significative, un test de Student doit être effectué pour déterminer si chaque variable indépendante est significative. Le test de signification individuelle de Student est présenté ci-dessous.

► **Test de signification individuelle de Student**

Pour tout paramètre β_i,

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

► **Statistique de test**

$$t = \frac{b_i}{s_{b_i}} \tag{13.15}$$

► **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 $n - p - 1$ degrés de liberté.

Dans la statistique de test, s_{b_i} correspond à l'estimation de l'écart type de b_i. La valeur de s_{b_i} est fournie par le logiciel.

Effectuons le test de Student dans le cadre du problème de régression de la société Butler. Le résultat de la programmation sous Minitab, reproduit à la figure 13.6, révèle que b_1 est égal à 0,061135, b_2 à 0,9234, s_{b_1} à 0,009888 et s_{b_2} à 0,2211. Ainsi, en utilisant l'équation (13.15), on obtient les valeurs suivantes pour les statistiques des tests d'hypothèses relatifs aux paramètres β_1 et β_2 :

$$t = 0{,}061135 / 0{,}009888 = 6{,}18$$

$$t = 0{,}9234 / 0{,}2211 = 4{,}18$$

Notez que ces deux valeurs t et les valeurs p correspondantes sont fournies par Minitab (cf. figure 13.6). Au seuil $\alpha = 0{,}01$, les valeurs p égales à 0,000 et 0,004 permettent de conclure au rejet des hypothèses $H_0 : \beta_1 = 0$ et $H_0 : \beta_2 = 0$. Par conséquent, les deux paramètres sont statistiquement significatifs. De même, la table 2 de l'annexe B indique qu'avec $n - p - 1 = 10 - 2 - 1 = 7$ degrés de liberté, la valeur critique est égale à $t_{0{,}005} = 3{,}499$. Avec $6{,}18 > 3{,}499$, on rejette l'hypothèse $H_0 : \beta_1 = 0$. De façon similaire, puisque $4{,}18 > 3{,}499$, on rejette également l'hypothèse $H_0 : \beta_2 = 0$.

13.5.3 Multi-colinéarité

Nous utilisons le terme « variables indépendantes » dans l'analyse de la régression pour parler des variables utilisées pour expliquer la valeur de la variable dépendante. Ce terme ne signifie pas que les variables indépendantes sont elles-mêmes indépendantes au sens statistique du terme. Au contraire, la plupart des variables indépendantes dans un problème de régression multiple sont plus ou moins corrélées les unes aux autres. Par exemple, dans l'exemple de la société de transport Butler impliquant deux variables indépendantes, le nombre de kilomètres parcourus et le nombre de livraisons effectuées, nous pouvons considérer le nombre de kilomètres parcourus comme une variable dépendante, expliquée par le nombre de livraisons effectuées. Il est alors possible de calculer le coefficient de corrélation de l'échantillon $r_{x_1x_2}$ pour déterminer dans quelle mesure ces deux variables sont liées. En appliquant ce raisonnement, on trouve $r_{x_1x_2} = 0{,}16$. Ainsi, les deux variables indépendantes sont, dans une certaine mesure, linéairement associées. En analyse de la régression multiple, la **multi-colinéarité** fait référence à la corrélation entre les variables indépendantes.

Pour approfondir les éventuels problèmes liés à la multi-colinéarité, considérons une variante de l'exemple de la société de transport Butler. Au lieu de considérer que x_2 correspond au nombre de livraisons, posons x_2 égal au nombre de litres de gasoil consommés. Clairement, x_1 (le nombre de kilomètres parcourus) et x_2 sont liés : le nombre de litres de gasoil consommés dépend du nombre de kilomètres parcourus. Par conséquent, nous devrions logiquement conclure que x_1 et x_2 sont des variables indépendantes fortement corrélées.

Supposez que nous obtenions l'équation $\hat{y} = b_0 + b_1x_1 + b_2x_2$ et que le test de Fisher révèle que la relation est significative. Supposez alors que nous effectuions un test

de Student sur β_1 pour déterminer si $\beta_1 \neq 0$, et que nous ne puissions rejeter $H_0 : \beta_1 = 0$. Ce résultat signifie-t-il que le temps de trajet n'est pas lié à la distance parcourue ? Pas nécessairement. Ce que cela signifie probablement, c'est qu'avec la présence de x_2 dans le modèle, x_1 ne contribue pas de façon significative à déterminer la valeur de y. Cette interprétation fait sens dans notre exemple : si nous connaissons la quantité de gasoil consommée, la connaissance du nombre de kilomètres parcourus n'apporte pas beaucoup d'informations complémentaires, utiles pour prévoir y. De même, un test de Student pourrait conduire à conclure que $\beta_2 = 0$, dans la mesure où la connaissance de la quantité de gasoil consommée n'apporte pas d'informations complémentaires significatives dans un modèle comprenant déjà le nombre de kilomètres parcourus.

Pour résumer, dans le test de signification individuelle de Student, la multi-colinéarité peut conduire à conclure qu'aucun des paramètres, pris individuellement, n'est significativement différent de zéro, alors que le test de signification globale de Fisher révèle une relation significative. Ce problème ne se pose pas lorsqu'il y a peu de corrélation entre les variables indépendantes.

Un coefficient de corrélation entre deux variables indépendantes supérieur à +0,70 ou inférieur à –0,70 indique l'existence de potentiels problèmes liés à la multi-colinéarité.

Les statisticiens ont développé plusieurs tests pour déterminer si l'ampleur de la multi-colinéarité pouvait poser problème. Selon le test de la règle de raison, la multi-colinéarité pose potentiellement problème si la valeur absolue du coefficient de corrélation de l'échantillon entre deux variables indépendantes est supérieure à 0,7. Les autres types de test sont plus avancés et vont au-delà de l'objet de cet ouvrage.

Lorsque les variables indépendantes sont fortement corrélées, il n'est pas possible de déterminer l'effet propre d'une variable indépendante particulière sur la variable dépendante.

Si possible, essayez de ne pas inclure dans le modèle des variables indépendantes fortement corrélées. En pratique, cependant, il est difficile de mettre en œuvre cette recommandation. Lorsque vous êtes en présence de multi-colinéarité, séparer l'impact individuel des variables indépendantes sur la variable dépendante est difficile.

REMARQUES

D'ordinaire, la multi-colinéarité n'affecte pas la procédure d'analyse de la régression ou l'interprétation des résultats. Toutefois, lorsque la multi-colinéarité est très prononcée – c'est-à-dire lorsque plusieurs variables indépendantes sont fortement corrélées – l'interprétation des résultats du test de Student peut s'avérer difficile. En plus du type de problème illustré dans cette section, une forte multi-colinéarité peut conduire à des estimations par les moindres carrés de signe opposé. En d'autres termes, lors de

simulations dans lesquelles les chercheurs créent un modèle de régression, estiment β_0, β_1, β_2, etc., il a été prouvé qu'en présence d'une forte multi-colinéarité, les estimations par les moindres carrés peuvent avoir un signe opposé à celui du paramètre estimé. Par exemple, β_2 peut être égal à +10 et b_2 estimé à – 2. En conséquence, peu de crédibilité doit être accordée aux coefficients individuels si on est en présence de multi-colinéarité.

EXERCICES

Méthode

19. Dans l'exercice 1, l'équation estimée de la régression suivante, fondée sur dix observations, était présentée.

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

De plus, $SCT = 6\ 724{,}125$, $SCreg = 6\ 216{,}375$, $s_{b_1} = 0{,}0813$ et $s_{b_2} = 0{,}0567$.

a) Calculer *MCreg* et *MCres*.

b) Calculer la statistique de test F et effectuer le test de Fisher. Utiliser $\alpha = 0{,}05$.

c) Effectuer le test de signification individuelle pour β_1.Utiliser $\alpha = 0{,}05$.

d) Effectuer le test de signification individuelle pour β_2.Utiliser $\alpha = 0{,}05$.

20. Référez-vous aux données de l'exercice 2. L'équation estimée de la régression associée à ces données est

$$\hat{y} = -18{,}37 + 2{,}01x_1 + 4{,}74x_2$$

$SCT = 15\ 182{,}9$, $SCreg = 14\ 052{,}2$, $s_{b_1} = 0{,}2471$ et $s_{b_2} = 0{,}9484$.

a) Tester l'existence d'une relation significative entre x_1, x_2 et y. Utiliser $\alpha = 0{,}05$.

b) β_1 est-il significatif ? Utiliser $\alpha = 0{,}05$.

c) β_2 est-il significatif ? Utiliser $\alpha = 0{,}05$.

21. L'équation estimée de la régression suivante a été développée pour un modèle à deux variables indépendantes.

$$\hat{y} = 40{,}7 + 8{,}63x_1 + 2{,}71x_2$$

La variable x_2 a été supprimée du modèle. L'application de la méthode des moindres carrés au modèle ne comprenant que x_1 comme variable indépendante fournit l'équation estimée de la régression suivante.

$$\hat{y} = 42{,}0 + 9{,}01x_1$$

a) Interpréter le coefficient associé à x_1 dans les deux modèles.

b) La multi-colinéarité peut-elle expliquer pourquoi le coefficient associé à x_1 diffère entre les deux modèles ? Si oui, comment ?

Applications

22. Dans l'exercice 4, l'équation estimée de la régression suivante, reliant les ventes au stock de marchandises et aux dépenses publicitaires, était donnée.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Les données utilisées pour développer ce modèle sont issues d'une enquête auprès de dix magasins. Pour ces données $SCT = 16\,000$ et $SCreg = 12\,000$.

a) Calculer *SCres*, *MCres* et *MCreg*.

b) Effectuer un test de Fisher avec $\alpha = 0{,}05$ pour déterminer l'existence d'une relation significative entre les variables.

23. Référez-vous à l'exercice 5.

a) Utiliser $\alpha = 0{,}01$ pour tester les hypothèses suivantes :

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : \beta_1 \text{ et/ou } \beta_2 \text{ n'est pas égal à zéro}$$

pour le modèle $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ où x_1 correspond aux dépenses publicitaires télévisées (en milliers de dollars) et x_2 aux dépenses publicitaires dans les journaux (en milliers de dollars).

b) Utiliser $\alpha = 0{,}05$ pour tester la significativité du paramètre β_1. La variable x_1 devrait-elle être retirée du modèle ?

c) Utiliser $\alpha = 0{,}05$ pour tester la significativité du paramètre β_2. La variable x_2 devrait-elle être retirée du modèle ?

24. La ligue nationale de football (NFL) enregistre différentes données sur les performances des individus et des équipes. Une partie des données indiquant le nombre moyen de yards gagnés par jeu offensif (OffPassYds/jeu), le nombre moyen de yards abandonnés par jeu défensif (DefYds/jeu) et le pourcentage de parties gagnées (% parties gagnées) au cours de la saison 2011 (cf. fichier en ligne NFL2011) est reprise ci-dessous (site Internet de ESPN, 3 novembre 2012).

Équipe	OffPassYds/jeu	DefYds/jeu	% parties gagnées
Arizona	222,9	355,1	50
Atlanta	262,0	333,6	62,5
Baltimore	213,9	288,9	75,0
.	.	.	.
.	.	.	.
.	.	.	.
St. Louis	179,4	358,4	12,5
Tampa Bay	228,1	394,4	25,0
Tennessee	245,2	355,1	56,3
Washington	235,8	339,8	31,3

a) Développer l'équation estimée de la régression qui peut être utilisée pour prévoir le pourcentage de parties gagnées étant donnés le nombre moyen de yards gagnés par jeu offensif et le nombre moyen de yards abandonnés par jeu défensif.

b) Utiliser le test de Fisher pour déterminer si la relation est globalement significative. Quelle est votre conclusion au seuil $\alpha = 0{,}05$?

c) Utiliser le test de Student pour déterminer si chaque variable indépendante est statistiquement significative. Quelle est votre conclusion au seuil $\alpha = 0{,}05$?

25. La liste Or 2012 de *Condé Nast Traveler* a fourni les évaluations des 20 meilleures croisières en bateau (site Internet de *Condé Nast Traveler*, 1er mars 2012). Les données reprises ci-dessous (cf. fichier en ligne Bateau) correspondent aux notes attribuées à chaque bateau de croisière, fondées sur les résultats de l'enquête annuelle Readers' Choice menée par *Condé Nast Traveler*. Chaque note représente le pourcentage de personnes interrogées qui ont évalué le bateau comme excellent ou très bon selon plusieurs critères comme l'itinéraire, les excursions sur le littoral et les repas. Une note globale est également reportée et utilisée pour classer les bateaux. Le premier bateau du classement, le Seabourn Odyssey, a obtenu une note globale de 94,4, et la note associée aux repas la plus élevée égale à 97,8.

Bateau	Note globale	Itinéraire	Excursions sur le littoral	Repas
Seabourn Odyssey	94,4	94,6	90,9	97,8
Seabourn Pride	93,0	96,7	84,2	96,7
National Geographie Endeavor	92,9	100,0	100,0	88,5
Seabourn Sojourn	91,3	88,6	94,8	97,1
Paul Gauguin	90,5	95,1	87,9	81,2
Seabourn Legend	90,3	92,5	82,1	98,8
Seabourn Spirit	90,2	96,0	86,3	92,0
Silver Explorer	89,9	92,6	92,6	88,9
Silver Spirit	89,4	94,7	85,9	90,8
Seven Seas Navigator	89,2	90,6	83,3	90,5
Silver Whisperer	89,2	90,9	82,0	88,6
National Geographic Explorer	89,1	93,1	93,1	89,7
Silver Cloud	88,7	92,6	78,3	91,3
Celebrity Xpedition	87,2	93,1	91,7	73,6
Silver Shadow	87,2	91,0	75,0	89,7
Silver Wind	86,6	94,4	78,1	91,6

Bateau	Note globale	Itinéraire	Excursions sur le littoral	Repas
SeaDream II	86,2	95,5	77,4	90,9
Wind Star	86,1	94,9	76,5	91,5
Wind Surf	86,1	92,1	72,3	89,3
Wind Spirit	85,2	93,5	77,4	91,9

a) Développer l'équation estimée de la régression qui permet de prévoir la note globale étant données les évaluations faites de l'itinéraire, des excursions et des repas.

b) Effectuer un test de Fisher pour déterminer si la relation est globalement significative. Quelle est votre conclusion au seuil $\alpha = 0,05$?

c) Effectuer un test de Student pour déterminer si chaque variable indépendante est statistiquement significative. Quelle est votre conclusion au seuil $\alpha = 0,05$?

d) Supprimer les variables indépendantes qui ne seraient pas significatives de l'équation estimée de la régression. Quelle équation estimée de la régression recommanderiez-vous ?

26. Dans l'exercice 10, des données (cf. fichier en ligne MLB) relatives aux valeurs de plusieurs statistiques sur les lancers pour un échantillon aléatoire de 20 lanceurs de la ligue américaine de la MLB ont été fournies (site Internet de la MLB, 1er mars 2012). À la question (c) de cet exercice, une équation estimée de la régression a été développée reliant le nombre moyen de coups sûrs par manche aux nombres moyens de strikeouts et de home runs par manche.

a) Effectuer un test de Fisher pour déterminer si la relation est globalement significative. Quelle est votre conclusion au seuil $\alpha = 0,05$?

b) Effectuer un test de Student pour déterminer si chaque variable indépendante est statistiquement significative. Quelle est votre conclusion au seuil $\alpha = 0,05$?

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

Les procédures d'estimation de la moyenne de y et de prévision d'une valeur de y dans le cadre d'une régression multiple sont similaires à celles employées dans le cadre d'une régression linéaire simple. Tout d'abord, rappelons qu'au chapitre 12, nous avons montré que l'estimation ponctuelle de la moyenne de y pour une valeur donnée de x était identique à l'estimation ponctuelle d'une valeur individuelle de y. Dans les deux cas, nous avons utilisé $\hat{y} = b_0 + b_1 x$ comme estimation ponctuelle.

La même procédure est utilisée pour une régression multiple. Nous substituons les valeurs données des variables indépendantes dans l'équation estimée de la régression et utilisons la valeur correspondante de $\hat{y}$ comme estimation ponctuelle. Supposez que

nous voulions, dans le cadre de l'exemple de la société de transport Butler, utiliser l'équation estimée de la régression impliquant x_1 (le nombre de kilomètres parcourus) et x_2 (le nombre de livraisons effectuées) pour construire deux estimations par intervalle :

3. Un *intervalle de confiance* du temps moyen de trajet pour tous les camions qui effectuent 100 km et deux livraisons
4. Un *intervalle de prévision* du temps de trajet d'un camion *spécifique* qui effectue 100 km et deux livraisons

En utilisant l'équation estimée de la régression $\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2$ avec $x_1 = 100$ et $x_2 = 2$, on obtient

$$\hat{y} = -0{,}869 + 0{,}0611(100) + 0{,}923(2) = 7{,}09$$

Par conséquent, l'estimation ponctuelle du temps de trajet dans les deux cas est d'environ 7 heures.

Pour développer des estimations par intervalle de la moyenne de y et d'une valeur individuelle de y, nous utilisons une procédure similaire à celle utilisée dans le cadre de l'analyse de la régression linéaire simple, avec une seule variable indépendante. Les formules requises vont au-delà de l'objet de cet ouvrage. Les logiciels fournissent souvent des intervalles de confiance dans le cadre de leur fonction d'analyse de la régression. Le tableau 13.4 contient les intervalles de confiance et de prévision à 95 % dans le cadre de l'exemple de la société Butler pour des valeurs particulières de x_1 et x_2, obtenus avec Minitab. Notez que l'intervalle de prévision est plus large que l'intervalle de confiance. Cet écart reflète le fait que, pour des valeurs données de x_1 et x_2, nous pouvons estimer le temps de trajet moyen pour tous les camions de façon plus précise que nous ne pouvons prévoir le temps de trajet d'un camion spécifique.

Tableau 13.4 ***Intervalles de confiance et de prévision à 95 % dans le cadre de l'exemple de la société Butler***

Valeur de x_1	Valeur de x_2	Intervalle de confiance		Intervalle de prévision	
		Limite inférieure	Limite supérieure	Limite inférieure	Limite supérieure
50	2	3,146	4,924	2,414	5,656
50	3	4,127	5,789	3,368	6,548
50	4	4,815	6,948	4,157	7,607
100	2	6,258	7,926	5,500	8,683
100	3	7,385	8,645	6,520	9,510
100	4	8,135	9,742	7,362	10,515

EXERCICES

Méthode

27. Dans l'exercice 1, l'équation estimée de la régression suivante, fondée sur dix observations, était présentée.

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

a) Développer une estimation ponctuelle de la moyenne de y lorsque $x_1 = 180$ et $x_2 = 310$.

b) Développer une estimation ponctuelle d'une valeur individuelle de y lorsque $x_1 = 180$ et $x_2 = 310$.

28. Référez-vous aux données de l'exercice 2. L'équation estimée de la régression associée à ces données est

$$\hat{y} = -18{,}4 + 2{,}01x_1 + 4{,}74x_2$$

a) Construire un intervalle de confiance à 95 % de la moyenne de y lorsque $x_1 = 45$ et $x_2 = 15$.

b) Construire un intervalle de prévision à 95 % pour y lorsque $x_1 = 45$ et $x_2 = 15$.

Applications

29. Dans l'exercice 5, le propriétaire de la société Showtime Movie Theaters utilisait l'analyse de la régression multiple pour prévoir le chiffre d'affaires (y) en fonction des dépenses publicitaires télévisées (x_1) et dans les journaux (x_2). L'équation estimée de la régression était

$$\hat{y} = 83{,}2 + 2{,}29x_1 + 1{,}30x_2$$

a) Quel est le chiffre d'affaires attendu lorsque 3 500 dollars sont dépensés en publicité télévisée ($x_1 = 3{,}5$) et 1 800 dollars en publicité dans les journaux ($x_2 = 1{,}8$) ?

b) Construire un intervalle de confiance à 95 % du chiffre d'affaires moyen associé aux dépenses publicitaires mentionnées à la question (a).

c) Construire un intervalle de prévision à 95 % du chiffre d'affaires d'une semaine particulière au cours de laquelle les dépenses publicitaires mentionnées à la question (a) ont été effectuées.

30. Dans l'exercice 24 (cf. fichier en ligne NFL), une équation estimée de la régression a été développée reliant le pourcentage de parties gagnées par une équipe de la NFL au cours de la saison 2011 (y) au nombre moyen de yards gagnés par jeu offensif (x_1) et au nombre moyen de yards abandonnés par jeu défensif (x_2) (site Internet de ESPN, 3 novembre 2012). Cette équation estimée de la régression était $\hat{y} = 60{,}5 + 0{,}319x_1 - 0{,}241x_2$.

a) Prédire le pourcentage de parties gagnées par une équipe particulière qui en moyenne gagne 225 yards par jeu offensif et abandonne en moyenne 300 yards par jeu défensif.

b) Construire un intervalle de confiance à 95 % pour le pourcentage moyen de parties gagnées pour toutes les équipes qui, en moyenne, gagnent 225 yards par jeu offensif et abandonnent en moyenne 300 yards par jeu défensif.

31. L'enquête en ligne sur les courtiers de l'Association Américaine des Investisseurs Individuels (AAII) interroge les membres de l'association sur leurs expériences avec des courtiers. On demande notamment aux membres d'évaluer le coût de la transaction et 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 de 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	Coût de la transaction	Vitesse	Satisfaction
Scottrade, Inc.	3,4	3,4	3,5
Charles Schwab	3,2	3,3	3,4
Fidelity Brokerage Services	3,1	3,4	3,9
TD Ameritrade	2,9	3,6	3,7
E*Trade Financial	2,9	3,2	2,9
(Non listé)	2,5	3,2	2,7
Vanguard Brokerage Services	2,6	3,8	2,8
USAA Brokerage Services	2,4	3,8	3,6
Thinkorswim	2,6	2,6	2,6
Wells Fargo Investments	2,3	2,7	2,3
Interactive Brokers	3,7	4,0	4,0
Zecco.com	2,5	2,5	2,5
Firstrade Securities	3,0	3,0	4,0
Bank of America Investment Services	4,0	1,0	2,0

a) Développer une équation estimée de la régression en utilisant le coût de la transaction et la vitesse d'exécution pour prévoir la satisfaction globale vis-à-vis du courtier.

b) Finger Lakes Investments a développé un nouveau système de transactions électroniques et souhaiterait prévoir la satisfaction globale des clients en supposant que ce nouveau système peut fournir des niveaux de satisfaction égaux à 3 en termes de coût de transaction et de vitesse d'exécution. Utiliser l'équation estimée de la régression développée à la question (a) pour prévoir le niveau de satisfaction globale des clients vis-à-vis de Finger Lakes Investments, si l'entreprise atteint ces niveaux de performance.

c) Construire un intervalle de confiance à 95 % de la note de satisfaction globale de tous les courtiers qui fournissent les mêmes niveaux de satisfaction de services que Finger Lakes Investments.

d) Construire un intervalle de prévision à 95 % de la note de satisfaction globale pour Finger Lakes Investments, en supposant que l'entreprise atteigne des niveaux de service égaux à 3 pour le coût de transaction et la vitesse d'exécution.

13.7 DES VARIABLES INDÉPENDANTES QUALITATIVES

Les variables indépendantes peuvent être qualitatives ou quantitatives.

Jusqu'à présent, les exemples considérés concernaient des variables indépendantes quantitatives telles que la population d'étudiants, la distance parcourue et le nombre de livraisons. Dans beaucoup de situations, cependant, nous devons travailler avec des **variables indépendantes qualitatives** telles que le sexe (homme ou femme), le mode de paiement (espèces, carte de crédit, chèque), etc. Le but de cette section est de montrer comment sont traitées les variables qualitatives dans l'analyse de la régression. Pour illustrer leur utilisation et leur interprétation, nous considérons un problème rencontré par les responsables de la société Johnson Filtration.

13.7.1 Un exemple : la société Johnson Filtration

La société Johnson Filtration offre des services de maintenance des systèmes de filtration d'eau dans le Sud de la Floride. Des clients souhaitant entretenir leurs systèmes de filtration d'eau, contactent la société Johnson. Pour estimer le temps et le coût du service offert, les responsables de la société Johnson souhaitent prévoir le temps de réparation nécessaire à chaque demande d'intervention. Dans ce contexte, le temps de réparation (en heures) correspond à la variable dépendante. Le temps de réparation est supposé lié à deux facteurs : le nombre de mois écoulés depuis la dernière

Tableau 13.5 ***Données associées à l'exemple de la société Johnson Filtration***

Demande d'intervention	Mois écoulés depuis la dernière intervention	Type de réparation	Durée de la réparation en heures
1	2	Électrique	2,9
2	6	Mécanique	3,0
3	8	Électrique	4,8
4	3	Mécanique	1,8
5	2	Électrique	2,9
6	7	Électrique	4,9
7	9	Mécanique	4,2
8	8	Mécanique	4,8
9	4	Électrique	4,4
10	6	Électrique	4,5

intervention et le type de problème nécessitant réparation (mécanique ou électrique). Les données relatives à un échantillon de dix demandes d'intervention sont présentées dans le tableau 13.5.

Soient y le temps de réparation en heures et x_1 le nombre de mois écoulés depuis la dernière intervention. Le modèle de régression utilisant x_1 pour prévoir y est

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

En utilisant Minitab pour estimer l'équation de la régression, nous obtenons les résultats présentés à la figure 13.7. L'équation estimée de la régression est

$$\hat{y} = 2{,}15 + 0{,}304 x_1 \tag{13.16}$$

Au seuil de signification de 0,05, la valeur p associée au test de Student (ou au test de Fisher), égale à 0,016, indique que le nombre de mois écoulés depuis la dernière intervention est significativement lié à la durée de la réparation. $R^2 = 53{,}4$ % indique que x_1 explique à lui seul 53,4 % de la variabilité de la durée des réparations.

Pour incorporer le type de réparation dans le modèle de régression, nous définissons la variable suivante :

$$x_2 = \begin{cases} 0 \text{ si la réparation est de type mécanique} \\ 1 \text{ si la réparation est de type électrique} \end{cases}$$

```
The regression equation is
Time = 2.15 + 0.304 Months

Predictor         Coef    SE Coef       T       p
Constant        2.1473     0.6050    3.55   0.008
Months          0.3041     0.1004    3.03   0.016

S = 0.7810    R-sq = 53.4%    R-sq(adj) = 47.6%

Analysis of Variance

SOURCE           DF       SS       MS       F       p
Regression        1   5.5960   5.5960    9.17   0.016
Residual Error    8   4.8800   0.6100
Total             9  10.4760
```

Figure 13.7 ***Output Minitab dans le cadre de l'exemple de la société Johnson Filtration, avec, pour variable indépendante, le nombre de mois écoulés depuis la dernière intervention***

Les noms des variables apparaissant dans l'output Minitab « Month » (mois) et « Time » (durée) ont été enregistrés en tant qu'intitulé des colonnes de la feuille de calcul Minitab. Ainsi, x_1 = Month et y = Time.

Dans l'analyse de la régression, x_2 est qualifiée de **variable muette** ou **variable indicatrice**. Grâce à cette variable muette, nous pouvons écrire le modèle de régression multiple comme suit

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Le tableau 13.6 (cf. fichier en ligne Johnson) correspond à l'ensemble de données révisé, incluant les valeurs de la variable muette. En utilisant Minitab pour estimer les paramètres du modèle et les données du tableau 13.6, nous obtenons l'équation estimée de la régression multiple suivante (cf. figure 13.8).

$$\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2 \qquad (13.17)$$

Au seuil de signification de 0,05, la valeur p égale à 0,01, associée au test de Fisher ($F = 21{,}36$), indique que la relation est significative. La partie de l'output (figure 13.8) relative au test de Student indique qu'à la fois, le nombre de mois écoulés depuis la dernière intervention (la valeur p est égale à 0,000) et le type de réparation (la valeur p est égale à 0,005) sont statistiquement significatifs. De plus, $R^2 = 85{,}9$ % et $R_a^2 = 81{,}9$ % indiquent que l'équation estimée de la régression explique une bonne part de la variabilité de la durée des réparations. Ainsi, l'équation (13.17) peut se révéler utile pour estimer le temps de réparation nécessaire pour répondre à différentes demandes.

13.7.2 Interpréter les paramètres

L'équation de régression multiple dans l'exemple de la société Johnson Filtration est

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (13.18)$$

Pour comprendre comment interpréter les paramètres β_0, β_1 et β_2 lorsqu'une variable qualitative est présente, considérons le cas où $x_2 = 0$ (réparation mécanique). En notant $E(y|\text{mécanique})$ l'espérance mathématique de la durée de réparation sachant que cette dernière est de type mécanique, nous obtenons

$$E(y|\text{mécanique}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \qquad (13.19)$$

De même, pour une réparation de type électrique ($x_2 = 1$), nous obtenons

$$E(y|\text{électrique}) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1 \qquad (13.20)$$

En comparant les équations (13.19) et (13.20), il apparaît que la durée de réparation est une fonction linéaire de x_1 à la fois pour des réparations mécaniques et électriques. La pente de ces deux équations est β_1, mais l'ordonnée à l'origine diffère. Elle est égale à β_0 dans l'équation (13.19) pour des réparations de type mécanique et à $(\beta_0 + \beta_2)$ dans l'équation (13.20) pour des réparations de type électrique. Ainsi, β_2 indique l'écart entre le temps moyen de réparation d'un problème électrique et le temps moyen de réparation d'un problème mécanique.

Si β_2 est positif, le temps moyen de réparation d'un problème électrique sera supérieur à celui d'un problème mécanique ; si β_2 est négatif le temps moyen de réparation

d'un problème électrique sera inférieur à celui d'un problème mécanique. Enfin, si $\beta_2 = 0$, il n'y a aucun écart entre la durée moyenne de réparation d'un problème électrique et d'un problème mécanique et la durée de réparation n'est pas liée à son type.

En utilisant l'équation estimée de la régression multiple $\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2$, nous constatons que 0,93 est l'estimation de β_0 et 1,26 l'estimation de β_2. Ainsi, lorsque $x_2 = 0$ (réparation mécanique),

$$\hat{y} = 0{,}93 + 0{,}388x_1 \tag{13.21}$$

et lorsque $x_2 = 1$ (réparation électrique),

$$\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26(1) = 2{,}19 + 0{,}388x_1 \tag{13.22}$$

L'utilisation d'une variable muette pour désigner le type de réparation fournit deux équations permettant de prévoir la durée des réparations ; l'une correspond aux réparations mécaniques, l'autre aux réparations électriques. De plus, avec $b_2 = 1{,}26$, nous savons qu'en général, les réparations électriques nécessitent 1,26 heure de plus que les réparations mécaniques.

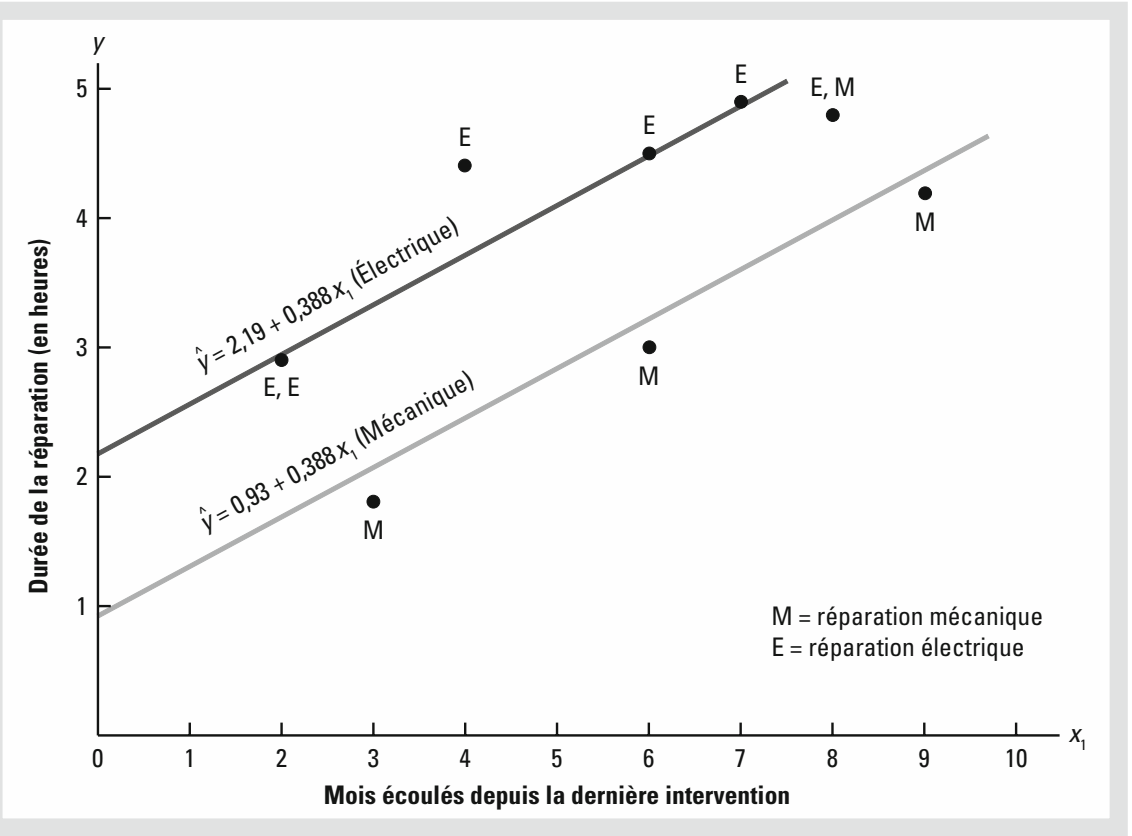

Figure 13.8 *Nuage de points des données de la société Johnson Filtration issues du tableau 13.6*

La figure 13.9 correspond au graphique des données de la société Johnson, présentées dans le tableau 13.6. La durée de réparation (en heures) est représentée sur l'axe vertical et le nombre de mois écoulés depuis la dernière intervention (x_1) est représenté sur l'axe horizontal. Un point correspondant à une réparation mécanique est indiqué par un M et un point correspondant à une réparation électrique est indiqué par un E. Les équations (13.21) et (13.22) sont représentées sur ce graphique pour illustrer graphiquement les deux équations qui peuvent être utilisées pour prévoir la durée d'une réparation, l'une correspondant à des réparations mécaniques, l'autre à des réparations électriques.

13.7.3 Des variables qualitatives plus complexes

Dans la mesure où la variable qualitative mentionnée dans l'exemple de la société Johnson Filtration a deux niveaux (mécanique ou électrique), définir une variable muette en indiquant une réparation de type mécanique par 0 et une réparation de type électrique par 1 est simple. Toutefois, lorsqu'une variable muette a plus de deux niveaux, il faut être attentif à la façon dont elle est définie et interprétée. Comme nous le verrons, si une variable qualitative a k niveaux, $k - 1$ variables muettes sont nécessaires, chacune prenant les valeurs 0 ou 1.

> Une variable qualitative à k niveaux doit être modélisée en utilisant $k - 1$ variables muettes. Il convient d'être attentif à la façon dont elles seront définies et interprétées.

Par exemple, supposons qu'un fabricant de photocopieuses ait réparti ses ventes dans un État particulier en trois régions : A, B et C. Les responsables souhaitent utiliser les techniques d'analyse de la régression pour prévoir le nombre de photocopieuses vendues par semaine. En prenant pour variable dépendante le nombre de photocopieuses vendues, ils considèrent plusieurs variables indépendantes (le nombre de vendeurs, les dépenses publicitaires, etc.). Supposons que les responsables pensent que la région de vente est également un facteur important pour prévoir le nombre de photocopieuses vendues. Puisque la région de vente est une variable qualitative à trois niveaux, A, B et C, nous avons besoin de 3 – 1 = 2 variables aléatoires pour représenter la région de vente. Chaque variable peut prendre la valeur 0 ou 1, comme indiqué ci-dessous.

$$x_1 = \begin{cases} 1 \text{ si la région de vente est B} \\ 0 \text{ sinon} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ si la région de vente est C} \\ 0 \text{ sinon} \end{cases}$$

Avec cette définition, nous obtenons les valeurs suivantes pour x_1 et x_2.

Région	x_1	x_2
A	0	0
B	1	0
C	0	1

Les observations relatives à la région A correspondent à $x_1 = 0$ et $x_2 = 0$; celles relatives à la région B correspondent à $x_1 = 1$ et $x_2 = 0$; celles relatives à la région C à $x_1 = 0$ et $x_2 = 1$.

L'équation de la régression reliant l'espérance mathématique du nombre de photocopieuses vendues, $E(y)$, aux variables muettes s'écrit :

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Pour aider à l'interprétation des paramètres β_0, β_1 et β_2, considérons les trois variantes suivantes de l'équation de la régression.

$$E(y \mid \text{région A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$
$$E(y \mid \text{région B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$
$$E(y \mid \text{région C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Ainsi, β_0 correspond à l'espérance mathématique du nombre de photocopieuses vendues dans la région A ; β_1 correspond à l'écart entre le nombre moyen d'unités vendues dans la région B et le nombre moyen d'unités vendues dans la région A ; et β_2à l'écart entre le nombre moyen d'unités vendues dans la région C et le nombre moyen d'unités vendues dans la région A.

Deux variables aléatoires étaient nécessaires dans la mesure où la région de vente est une variable qualitative à trois niveaux. Le fait que $x_1 = 0$ et $x_2 = 0$ indique la région A, $x_1 = 1$ et $x_2 = 0$ la région B et $x_1 = 0$ et $x_2 = 1$ la région C est arbitraire. Par exemple, nous aurions pu choisir d'indiquer la région A par $x_1 = 1$ et $x_2 = 0$, la région B par $x_1 = 0$ et $x_2 = 0$ et la région C par $x_1 = 0$ et $x_2 = 1$. Dans ce cas, β_1 correspondrait à l'écart entre le nombre moyen d'unités vendues dans les régions A et B ; et β_2à l'écart entre le nombre moyen d'unités vendues dans les régions C et B.

Le point important à retenir est que lorsqu'une variable qualitative a k niveaux, $k - 1$ variables muettes sont nécessaires dans le modèle de régression multiple. Ainsi, si une quatrième région D était ajoutée dans l'exemple précédent, trois variables muettes seraient nécessaires pour effectuer l'analyse. Elles pourraient éventuellement être codées de la façon suivante.

$$x_1 = \begin{cases} 1 \text{ si la région de vente est B} \\ 0 \text{ sinon} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ si la région de vente est C} \\ 0 \text{ sinon} \end{cases}$$

$$x_3 = \begin{cases} 1 \text{ si la région de vente est D} \\ 0 \text{ sinon} \end{cases}$$

EXERCICES

Méthode

32. Considérer l'étude d'une régression impliquant une variable dépendante y, une variable indépendante quantitative x_1 et une variable indépendante qualitative à deux niveaux (niveau 1 et niveau 2).

a) Écrire l'équation de la régression multiple reliant x_1 et la variable qualitative à y.

b) Quelle est l'espérance mathématique de y correspondant au niveau 1 de la variable qualitative ?

c) Quelle est l'espérance mathématique de y correspondant au niveau 2 de la variable qualitative ?

d) Interpréter les paramètres de votre équation de régression.

33. Considérer l'étude d'une régression impliquant une variable dépendante y, une variable indépendante quantitative x_1 et une variable indépendante qualitative à trois niveaux (niveau 1, niveau 2 et niveau 3).

a) Combien de variables muettes sont nécessaires pour représenter la variable qualitative ?

b) Écrire l'équation de la régression multiple reliant x_1 et la variable qualitative à y.

c) Interpréter les paramètres de votre équation de régression.

Applications

34. Des responsables ont proposé le modèle de régression suivant pour prévoir les ventes d'un fast-food.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

où y correspond aux ventes (en milliers de dollars), x_1 correspond au nombre de concurrents dans un rayon d'un kilomètre, x_2 à la population présente dans un rayon d'un kilomètre (en milliers) et $x_3 = \begin{cases} 1 \text{ si un service de drive-in est proposé} \\ 0 \text{ sinon} \end{cases}$.

L'équation estimée de la régression suivante a été développée à partir d'un échantillon de 20 fast-foods.

$$\hat{y} = 10{,}1 - 4{,}2x_1 + 6{,}8x_2 + 15{,}3x_3$$

a) Quel est le montant espéré des ventes attribuables à la présence d'un service de drive-in ?

b) Prévoir les ventes d'un fast-food implanté dans une zone comprenant deux concurrents et une population de 8 000 personnes dans un rayon d'un kilomètre, ne proposant pas de service de drive-in.

c) Prévoir les ventes d'un fast-food implanté dans une zone comprenant un seul concurrent et une population de 3 000 personnes dans un rayon d'un kilomètre, proposant un service de drive-in.

35. Référez-vous au problème de la société Johnson Filtration introduit dans cette section. Supposez qu'en plus de l'information concernant le nombre de mois écoulés depuis la dernière intervention et le type de panne (mécanique ou électrique), les responsables obtiennent le nom du réparateur. Les données révisées sont présentées ci-dessous (cf. fichier en ligne Réparation).

Durée de la réparation en heures	Mois écoulés depuis la dernière intervention	Type de réparation	Réparateur
2,9	2	Électrique	Dave Newton
3,0	6	Mécanique	Dave Newton
4,8	8	Électrique	Bob Jones
1,8	3	Mécanique	Dave Newton
2,9	2	Électrique	Dave Newton
4,9	7	Électrique	Bob Jones
4,2	9	Mécanique	Bob Jones
4,8	8	Mécanique	Bob Jones
4,4	4	Électrique	Bob Jones
4,5	6	Électrique	Dave Newton

a) Ignorer pour le moment le nombre de mois écoulés depuis la dernière intervention (x_1) et le réparateur. Développer l'équation estimée de la régression linéaire simple pour prévoir la durée de la réparation (y) en fonction du type de réparation (x_2). Pour mémoire, $x_2 = 0$ si la réparation est de type mécanique et $x_2 = 1$ si la réparation est de type électrique.

b) L'équation développée à la question (a) est-elle bien adaptée aux données observées ? Expliquer.

c) Ignorer pour le moment le nombre de mois écoulés depuis la dernière intervention et le type de réparation effectuée. Développer l'équation estimée de la régression linéaire simple pour prévoir la durée de la réparation (y) en fonction du réparateur. Si le réparateur est Bob Jones, $x_3 = 0$; si le réparateur est Dave Newton, $x_3 = 1$.

d) L'équation développée à la question (c) est-elle bien adaptée aux données observées ? Expliquer.

36. Ce problème est une extension de l'exercice 35.

a) Développer l'équation estimée de la régression pour prévoir le temps de réparation étant donnés le nombre de mois écoulés depuis la dernière intervention, le type de réparation et le réparateur.

b) Au seuil de signification de 0,05, tester l'existence d'une relation significative entre les variables indépendantes et la variable dépendante de la question (a).

c) L'ajout de la variable indépendante x_3, le réparateur, est-il statistiquement significatif ? Utiliser $\alpha = 0{,}05$. Quelle explication pouvez-vous apporter aux résultats observés ?

37. L'enquête de satisfaction des clients dans les restaurants menée par le magazine *Consumer Reports* est basée sur 148 499 visites dans des chaînes de restaurants (site Internet de Consumer Reports, 11 février 2009). Supposez que les données suivantes (cf. fichier

en ligne Restaurants) sont représentatives des résultats de l'enquête. La variable Type indique si le restaurant est un restaurant italien ou un restaurant de poisson/grill. Le prix indique le montant moyen payé par personne pour un repas et les boissons diminué du pourboire. La note reflète la satisfaction globale des clients, des valeurs plus élevées reflétant une satisfaction globale plus importante. Une note de 80 est considérée comme très satisfaisante.

Restaurant	Type	Prix ($)	Note
Bertucci's	Italien	16	77
Black Angus Steackhouse	Poisson/Grill	24	79
Bonefish Grill	Poisson/Grill	26	85
Bravo ! Cucina Italiana	Italien	18	84
Buca di Beppo	Italien	17	81
Bugaboo Creek Steak House	Poisson/Grill	18	77
Carrabba's Italian Grill	Italien	23	86
Charlie Brown's Steakhouse	Poisson/Grill	17	75
Il Fornaio	Italien	28	83
Joe's Crab Shack	Poisson/Grill	15	71
Johnny Carino's Italian	Italien	17	81
Lone Star Steakhouse & Saloon	Poisson/Grill	17	76
LongHorn Steakhouse	Poisson/Grill	19	81
Maggiano's Little Italy	Italien	22	83
McGrath's Fish House	Poisson/Grill	16	81
Olive Garden	Italien	19	81
Outback Steakhouse	Poisson/Grill	20	80
Red Lobster	Poisson/Grill	18	78
Romano's Macaroni Grill	Italien	18	82
The Old Spaghetti Factory	Italien	12	79
Uno Chicago Grill	Italien	16	76

a) Développer l'équation estimée de la régression qui permet de montrer la relation entre la satisfaction globale des clients et le prix moyen du repas.

b) Au seuil de signification de 0,05, tester si l'équation estimée de la régression développée à la question (a) indique une relation significative entre la satisfaction globale des clients et le prix moyen du repas.

c) Construire une variable muette représentant le type de restaurant (italien ou de poisson/grill).

d) Développer l'équation estimée de la régression qui montre comment la satisfaction globale des clients est liée au prix moyen du repas et au type de restaurant.

e) Le type de restaurant est-il un facteur significatif expliquant la satisfaction globale des clients ?

f) Estimer la satisfaction globale d'un client déjeunant dans un restaurant de poisson/grill pour 20 dollars. Quel serait l'écart entre cette note et celle obtenue si le restaurant était un italien ?

38. Une étude menée pendant 10 ans par l'association américaine Heart a fourni des données sur l'impact de l'âge, de la pression artérielle et du fait de fumer sur le risque de faire un arrêt cardiaque. Supposez que les données suivantes (cf. fichier en ligne Arrêt cardiaque) soient une partie de cette étude. Le risque d'arrêt cardiaque est interprété comme la probabilité (multipliée par 100) que le patient ait une attaque au cours des dix prochaines années. Pour la variable « fumeur », définir une variable muette (1 indiquant un fumeur, 0 un non-fumeur).

Risque	Âge	Pression artérielle	Fumeur
12	57	152	Non
24	67	163	Non
13	58	155	Non
56	86	177	Oui
28	59	196	Non
51	76	189	Oui
18	56	155	Oui
31	78	120	Non
37	80	135	Oui
15	78	98	Non
22	71	152	Non
36	70	173	Oui
15	67	135	Oui
48	77	209	Oui
15	60	199	Non
36	82	119	Oui
8	66	166	Non
34	80	125	Oui
3	62	117	Non
37	59	207	Oui

a) Estimer l'équation de la régression reliant le risque d'une attaque à l'âge de la personne, sa pression artérielle et le fait que cette personne fume.

b) Le fait de fumer est-il un facteur significatif expliquant le risque d'une attaque ? Expliquer. Utiliser $\alpha = 0{,}05$.

c) Quelle est la probabilité que Art Speen, âgé de 68 ans, fumeur, dont la pression artérielle s'élève à 175, ait une attaque au cours des dix prochaines années ? Que pourrait recommander le médecin à son patient ?

RÉSUMÉ

Dans ce chapitre, nous avons introduit l'analyse de la régression multiple en tant qu'extension de l'analyse de la régression linéaire simple présentée au chapitre 12. L'analyse de la régression multiple nous permet de comprendre comment une variable dépendante est liée à au moins deux variables indépendantes. L'équation de régression

multiple $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$ indique que l'espérance mathématique ou la moyenne de la variable dépendante y est reliée aux valeurs des variables indépendantes $x_1, x_2, \ldots x_p$. Des données d'échantillon et la méthode des moindres carrés permettent d'estimer l'équation de la régression multiple $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$.

En effet, $b_0, b_1, b_2, \ldots, b_p$ sont des statistiques d'échantillon utilisées pour estimer les paramètres inconnus du modèle $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Les résultats fournis par les logiciels statistiques ont été utilisés à travers l'ensemble de ce chapitre, dans la mesure où il s'agit du seul moyen réaliste d'effectuer les calculs numériques nécessaires à l'analyse d'une régression multiple.

Le coefficient de détermination multiple a été présenté en tant que mesure de l'adéquation de l'équation estimée de la régression aux données de l'échantillon. Il détermine la proportion de la variabilité de y expliquée par l'équation estimée de la régression. Le coefficient de détermination multiple ajusté est une mesure similaire de l'adéquation de l'équation estimée de la régression, mais tenant compte du nombre de variables indépendantes et ainsi évitant de surestimer l'impact de l'ajout de variables indépendantes supplémentaires dans le modèle.

Les tests de Fisher et de Student ont été présentés en tant que moyens de déterminer statistiquement si la relation entre les variables est significative. Le test de Fisher permet de déterminer s'il y a une relation globalement significative entre la variable dépendante et l'ensemble des variables indépendantes. Le test de Student permet de déterminer s'il existe une relation significative entre la variable dépendante et une variable indépendante, étant données les autres variables indépendantes du modèle. La corrélation entre les variables indépendantes, dite multi-colinéarité, a été évoquée.

Le chapitre conclut sur l'utilisation des variables muettes en tant que moyen d'incorporer des variables indépendantes qualitatives dans l'analyse de la régression multiple.

GLOSSAIRE

Analyse de la régression multiple. Analyse de la régression impliquant plusieurs variables indépendantes.

Modèle de régression multiple. Équation qui décrit la relation entre la variable dépendante y et les variables indépendantes $x_1, x_2, \ldots, x_p$ et le terme d'erreur ε.

Équation de régression multiple. Équation qui décrit comment la moyenne de la variable dépendante est liée aux variables indépendantes ; $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.

Équation estimée de la régression multiple. Estimation de l'équation de régression multiple basée sur les données d'un échantillon et la méthode des moindres carrés ; $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$.

Méthode des moindres carrés. Procédure utilisée pour estimer l'équation de la régression. L'objectif est de minimiser la somme des résidus au carré (les écarts entre les valeurs observées de la variable dépendante y_i et ses valeurs estimées $\hat{y}_i$).

Coefficient de détermination multiple. Mesure de l'adéquation de l'équation estimée de la régression multiple. Il peut être interprété comme la part de la variation de la variable dépendante expliquée par l'équation estimée de la régression.

COEFFICIENT DE DÉTERMINATION MULTIPLE AJUSTÉ. Mesure de l'adéquation de l'équation estimée de la régression multiple, ajustée en fonction du nombre de variables indépendantes contenues dans le modèle, de façon à éviter de surestimer l'impact de l'ajout de variables indépendantes supplémentaires.

MULTI-COLINÉARITÉ. Terme utilisé pour décrire la corrélation entre les variables indépendantes.

VARIABLE INDÉPENDANTE QUALITATIVE. Variable indépendante dont les données sont qualitatives.

VARIABLE MUETTE. Variable utilisée pour modéliser l'impact de variables indépendantes qualitatives. Une variable muette ne peut prendre que les valeurs 0 ou 1.

FORMULES CLÉ

Modèle de régression multiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{13.1}$$

Équation de la régression multiple

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{13.2}$$

Équation estimée de la régression multiple

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \tag{13.3}$$

Critère des moindres carrés

$$\min \sum (y_i - \hat{y}_i)^2 \tag{13.4}$$

Relation entre SCT, SCreg et SCres

$$SCT = SCreg + SCres \tag{13.7}$$

Coefficient de détermination multiple

$$R^2 = SCreg / SCT \tag{13.8}$$

Coefficient de détermination multiple ajusté

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{13.9}$$

Moyenne des carrés de la régression

$$MCreg = \frac{SCreg}{p} \tag{13.12}$$

Moyenne des carrés des résidus

$$MCres = \frac{SCres}{n-p-1} \tag{13.13}$$

Statistique de test de Fisher

$$F = \frac{MCreg}{MCres} \tag{13.14}$$

Statistique de test de Student

$$t = \frac{b_i}{s_{b_i}} \tag{13.15}$$

EXERCICES SUPPLÉMENTAIRES

39. Le bureau des admissions de l'Université de Clearwater a développé l'équation estimée de la régression suivante, reliant la note moyenne obtenue à l'examen de fin d'année d'un étudiant à sa note en mathématique et sa moyenne au bac.

$$\hat{y} = -1{,}41 + 0{,}0235x_1 + 0{,}00486x_2$$

où x_1 correspond à la note moyenne obtenue au bac, x_2 à la note obtenue en mathématique et y à la note moyenne obtenue à l'examen de fin d'année.

a) Interpréter les coefficients de cette équation estimée de la régression.

b) Estimer la note moyenne obtenue à l'examen de fin d'année d'un étudiant qui a obtenu une note de 84 au bac et une note de 540 au test de mathématique.

40. Le directeur du personnel de la société Electronic Associates a développé l'équation de la régression suivante, reliant la note obtenue par un employé à un test de satisfaction professionnelle à son ancienneté et à son indice salarial.

$$\hat{y} = 14{,}4 - 8{,}69x_1 + 13{,}5x_2$$

où x_1 correspond à l'ancienneté (en années), x_2 à l'indice salarial et y à la note obtenue au test de satisfaction professionnelle (des notes élevées traduisent une plus grande satisfaction professionnelle).

a) Interpréter les coefficients de cette équation estimée de la régression.

b) Estimer la note qu'obtiendrait un employé qui a 4 années d'ancienneté et qui gagne 6,50 dollars de l'heure, au test de satisfaction professionnelle.

41. Une partie des résultats obtenus grâce à un logiciel dans le cadre de l'analyse d'une régression est présentée ci-dessous.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor       Coef   SE Coef       T
Constant      ______     2.667   _____
X1            ______     2.105   _____
X2            ______     0.613   _____

S = 3.335     R-sq = 92.3%    R-sq(adj) = ______%

Analysis of Variance

SOURCE             DF        SS       MS        F
Regression     ______      1612   ______   _____
Residual Error     12    ______   ______
Total          ______    ______
```

a) Compléter la feuille de résultats.

b) Effectuer le test de Fisher et tester au seuil $\alpha = 0{,}05$ l'existence d'une relation significative.

c) Utiliser le test de Student pour tester au seuil $\alpha = 0{,}05$ les hypothèses $H_0 : \beta_1 = 0$ et $H_0 : \beta_2 = 0$.

d) Calculer R_a^2.

42. Reprendre l'exercice 39. Le bureau des admissions de l'Université de Clearwater a développé l'équation estimée de la régression suivante, reliant la note moyenne obtenue à l'examen de fin d'année d'un étudiant à sa note en mathématique et sa moyenne au bac.

$$\hat{y} = -1{,}41 + 0{,}0235x_1 + 0{,}00486x_2$$

où x_1 correspond à la note moyenne obtenue au bac, x_2 à la note obtenue en mathématique et y à la note moyenne obtenue à l'examen de fin d'année.

Une partie des résultats obtenus grâce à Minitab dans le cadre de cette analyse est présentée ci-dessous.

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor              Coef        SE Coef            T
Constant            -1.4053         0.4848       _____
X1                 0.023467       0.008666       _____
X2                   ______       0.001077       _____

S = 0.1298      R-sq = ______%      R-sq(adj) =  ______%
```

```
Analysis of Variance

SOURCE                DF          SS          MS          F
Regression         ______     1.76209     ______     ______
Residual Error     ______     _______     ______
Total                  9     1.88000
```

a) Compléter l'output Minitab.

b) Effectuer le test de Fisher et tester au seuil $\alpha = 0{,}05$ l'existence d'une relation significative.

c) Utiliser le test de Student pour tester au seuil $\alpha = 0{,}05$ les hypothèses $H_0 : \beta_1 = 0$ et $H_0 : \beta_2 = 0$.

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

43. Reprendre l'exercice 40. Le directeur du personnel de la société Electronic Associates a développé l'équation de la régression suivante, reliant la note obtenue par un employé à un test de satisfaction professionnelle à son ancienneté et à son indice salarial.

$$\hat{y} = 14{,}4 - 8{,}69x_1 + 13{,}5x_2$$

où x_1 correspond à l'ancienneté (en années), x_2 à l'indice salarial et y à la note obtenue au test de satisfaction professionnelle (des notes élevées traduisent une plus grande satisfaction professionnelle).

Une partie des résultats obtenus grâce à Minitab dans le cadre de cette analyse est présentée ci-dessous.

```
The regression equation is
Y = - 1.41 + .0235 X1 + .00486 X2

Predictor              Coef        SE Coef           T
Constant            - 1.4053         0.4848     ______
X1                  0.023467       0.008666     ______
X2                    ______       0.001077     ______

S = 0.1298      R - sq = ______%      R - sq (adj) =  ______%

Analysis of Variance

SOURCE                DF          SS          MS          F
Regression         ______     1.76209     ______     ______
Residual Error     ______     _______     ______
Total                  9     1.88000
```

a) Compléter l'output Minitab.

b) Effectuer le test de Fisher et tester au seuil $\alpha = 0{,}05$ l'existence d'une relation significative.

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

d) Utiliser le test de Student pour tester au seuil $\alpha = 0{,}05$ les hypothèses $H_0 : \beta_1 = 0$ et $H_0 : \beta_2 = 0$.

44. Tire Rack, le distributeur en ligne leader aux États-Unis de pneus et de roues, mène de nombreux tests pour fournir à ses clients les produits adaptés à leur véhicule, à leur style de conduite et aux conditions de conduite auxquelles ils font face. De plus, Tire Rack actualise régulièrement une enquête indépendante auprès des consommateurs pour que les automobilistes s'aident mutuellement en partageant leurs expériences. Les données suivantes (cf. fichier en ligne TireRack) indiquent les notes (sur une échelle allant de 1 à 10, 10 étant la meilleure note) de performance de 18 pneus été (site Internet de Tire Rack, 3 février 2009). La variable Direction évalue la réactivité des pneus à des changements de direction, la variable Tenue évalue la tenue de route des pneus et la variable Rachat évalue la satisfaction globale de l'automobiliste et son désir de racheter le même pneu à l'avenir.

Pneu	Direction	Tenue	Rachat
Goodyear Assurance Triple Tred	8,9	8,5	8,1
Michelin HydroEdge	8,9	9,0	8,3
Michelin Harmony	8,3	8,8	8,2
Dunlop SP60	8,2	8,5	7,9
Goodyear Assurance ComforTred	7,9	7,7	7,1
Yokohama Y372	8,4	8,2	8,9
Yokohama Aegis LS4	7,9	7,0	7,1
Kumbo Power Star 758	7,9	7,9	8,3
Goodyear Assurance	7,6	5,8	4,5
Hankook H406	7,8	6,8	6,2
Michelin Energy LX4	7,4	5,7	4,8
Michelin MX4	7,0	6,5	5,3
Michelin Symmetry	6,9	5,7	4,2
Kumbo 722	7,2	6,6	5,0
Dunlop SP40 A/S	6,2	4,2	3,4
Bridgestone Insignia SE200	5,7	5,5	3,6
Goddyear Integrity	5,7	5,4	2,9
Dunlop SP20 FE	5,7	5,0	3,3

a) Estimer l'équation de la régression qui peut être utilisée pour prévoir l'évaluation globale (rachat) étant donnée la note attribuée à la variable Direction. Au seuil de 0,05, tester l'existence d'une relation significative.

b) L'équation estimée de la régression développée à la question (a) est-elle bien adaptée aux données ? Expliquer.

c) Développer l'équation estimée de la régression qui permet de prévoir la note de satisfaction globale (rachat) étant données les notes attribuées aux variables Direction et Tenue.

d) L'ajout de la variable indépendante Tenue est-elle utile ? Utiliser un seuil de signification de 0,05.

45. Le *Guide 2012 d'économie de l'essence* publié par le département américain à l'énergie et l'agence américaine de protection de l'environnement fournit des données sur la consommation d'essence des modèles 2012 de voitures et camions (site Internet du département de l'énergie, 16 avril 2012). Une partie des données relatives à 309 voitures est contenue dans le fichier en ligne intitulé Économie d'essence 2012. La colonne intitulée Fabricant indique le nom de l'entreprise qui a fabriqué la voiture ; la colonne intitulée Puissance indique le rapport volumétrique du moteur (en litres) ; la colonne intitulée Type de carburant indique si la voiture consomme de l'essence ordinaire (O) ou sans plomb (SP) ; la colonne intitulée Traction indique si la voiture est une traction avant (AV), une traction (AR) ou une quatre roues motrices (4R) et la colonne Consommation sur autoroute indique la consommation du véhicule en miles par gallon sur autoroute.

a) Développer une équation estimée de la régression permettant de prévoir la consommation sur autoroute étant donnée la puissance du moteur. Tester la significativité de la relation au seuil $\alpha = 0{,}05$.

b) Considérer l'ajout de la variable muette « Carburant SP » égale à 1 si la voiture consomme de l'essence sans plomb, 0 sinon. Développer l'équation estimée de la régression permettant de prévoir la consommation de carburant sur autoroute étant données la puissance du moteur et la variable muette « Carburant SP ».

c) Utiliser le seuil $\alpha = 0{,}05$ pour déterminer si l'ajout de la variable muette est significatif.

d) Considérez l'ajout des variables muettes AV et AR. La variable AV est égale 1 si la voiture est une traction avant, 0 sinon ; AR est égale à 1 si la voiture est une traction arrière, 0 sinon. Ainsi, pour une voiture quatre roues motrices, à la fois AV et AR sont égales à 0. Développer l'équation estimée de la régression permettant de prévoir la consommation de carburant sur autoroute étant données la puissance du moteur et les variables muettes « Carburant SP », « AR » et « AV ».

e) Pour l'équation estimée de la régression développée à la question (d), tester la significativité globale de la relation et la significativité individuelle des variables au seuil de 0,05.

46. Une partie de l'ensemble de données contenant les informations sur 45 fonds mutuels qui appartiennent au classement Morningstar Funds 500 de 2008 est fournie ci-dessous. L'ensemble de données complet est disponible en ligne dans le fichier intitulé Fonds Mutuels. L'ensemble de données contient les cinq variables suivantes :

Type : le fonds peut être constitué d'actions domestiques (D), internationales (I) ou d'actions à revenus fixes (F).

Valeur nette de l'actif (en dollars) : correspond au prix de clôture du cours de l'action au 31 décembre 2007.

Rendement moyen sur 5 ans (en pourcentage) : correspond au rendement annuel moyen du fonds au cours des 5 dernières années.

Ratio des dépenses (en pourcentage) : correspond au pourcentage d'actifs déduit couvrant les dépenses annuelles de fonctionnement du fonds.

Classement Morningstar : correspond à l'évaluation du risque du fonds faite par Morningstar, sur une échelle allant de 1 à 5 étoiles.

Nom du fonds	Type de fonds	Valeur nette de l'actif ($)	Rendement moyen sur 5 ans (%)	Ratio des dépenses (%)	Classement Morningstar (nombre d'étoiles)
Amer Cent Inc & Growth Inv	D	28,88	12,39	0,67	2
American Century Intl. Disc	I	14,37	30,53	1,41	3
American Century Tax-free Bond	F	10,73	3,34	0,49	4
American Century Ultra	D	24,94	10,88	0,99	3
Ariel	D	46,39	11,32	1,03	2
Artisan Intl. Val	I	25,52	24,95	1,23	3
Artisan Small Cap	D	16,92	15,67	1,18	3
Baron Asset	D	50,97	16,77	1,31	5
Brandwine	D	36,58	18,14	1,08	4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

a) Estimer l'équation de la régression qui peut être utilisée pour prévoir le rendement moyen sur 5 ans étant donné le type de fonds. Au seuil de 0,05, tester l'existence d'une relation significative.

b) L'équation estimée de la régression développée à la question (a) est-elle bien adaptée aux données ? Expliquer.

c) Estimer l'équation de la régression qui peut être utilisée pour prévoir le rendement moyen sur 5 ans étant donnés le type de fonds, la valeur nette de l'actif et le ratio des dépenses. Au seuil de 0,05, tester l'existence d'une relation significative. Pensez-vous que certaines variables devraient être retirées du modèle de régression ? Expliquer.

d) Le classement Morningstar est une variable qualitative. Puisque l'ensemble de données ne contient que des fonds qui ont entre 2 et 5 étoiles (4 rangs), utiliser les variables muettes suivantes : Rang-3 = 1 si le fonds a 3 étoiles, 0 sinon ; Rang-4 = 1 si le fonds a 4 étoiles, 0 sinon ; Rang-5 = 1 si le fonds a 5 étoiles, 0 sinon. Estimer l'équation de la régression qui peut être utilisée pour prévoir le rendement moyen sur 5 ans étant donnés le type de fonds, le ratio des dépenses et le classement Morningstar. En utilisant $\alpha = 0,05$, retirer du modèle toute variable indépendante qui n'est pas significative.

e) Utiliser l'équation estimée de la régression développée à la question (d) pour estimer le rendement moyen sur 5 ans d'un fonds domestique dont le ratio de dépenses est de 1,05 % et qui est classé 3 étoiles par Morningstar.

47. Le magazine *Fortune* publie une enquête annuelle des meilleures sociétés dans lesquelles travailler. Les données contenues dans le fichier en ligne Fortune Best reprend une partie des données pour un échantillon aléatoire de 30 sociétés appartenant au top 100 de cette liste en 2012 (*Fortune*, 6 février 2012). La colonne intitulée Rang indique le rang de la société dans le top 100 ; la colonne intitulée Taille indique si la société est une petite société, une société de taille moyenne ou une grande société ; la colonne intitulée Salariés (en milliers de dollars) indique le salaire annuel moyen des employés à temps complet, arrondi au milliers de dollars le plus proche ; et la colonne intitulée À l'heure (en milliers de dollars) indique le salaire annuel moyen des employés payés à l'heure, arrondi au millier de dollars le plus proche. *Fortune* définit les grandes sociétés comme celles ayant plus de 10 000 employés, les sociétés moyennes comme celles dont le nombre d'employés est compris entre 2 500 et 10 000 et les petites sociétés comme celles qui ont moins de 2 500 employés.

Rang	Société	Taille	Salariés (en milliers de dollars)	À l'heure (en milliers de dollars)
4	Wegmans Food Markets	Grande	56	29
6	NetApp	Moyenne	143	76
7	Camden Property Trust	Petite	71	37
8	Recreational Equipment (REI)	Grande	103	28
10	Quicken Loans	Moyenne	78	54
11	Zappos.com	Moyenne	48	25
12	Mercedes-Benz USA	Petite	118	50
20	USAA	Grande	96	47
22	The Container Store	Moyenne	71	45
25	Ultimate Software	Petite	166	56
37	Plante Moran	Petite	73	45
42	Baptist Health South Florida	Grande	126	80
50	World Wide Technology	Petite	129	31
53	Methodist Hospital	Grande	100	83
58	Perkins Coie	Petite	189	63
60	American Express	Grande	114	35
64	TDIndustries	Petite	93	47
66	QuikTrip	Grande	69	44
72	EOG Resources	Petite	189	81
75	FactSet Research Systems	Petite	103	51
80	Stryker	Grande	71	43
81	SRC	Petite	84	33
84	Booz Allen Hamilton	Grande	105	77
91	CarMax	Grande	57	34
93	GoDaddy.com	Moyenne	105	71
94	KPMG	Grande	79	59
95	Navy Federal Credit Union	Moyenne	77	39

Rang	Société	Taille	Salariés (en milliers de dollars)	À l'heure (en milliers de dollars)
97	Schweitzer Engineering Labs	Petite	99	28
99	Darden Restaurants	Grande	57	24
100	Intercontinental Hotels Group	Grande	63	26

a) Utiliser ces données pour estimer une équation de régression qui pourrait être utilisée pour prévoir le salaire annuel moyen des employés salariés à temps complet étant donné le salaire annuel moyen des employés à l'heure.

b) Utiliser $\alpha = 0{,}05$ pour tester la significativité globale de la relation.

c) Pour prendre en compte l'effet « taille », une variable qualitative à trois niveaux, nous avons utilisé deux variables muettes : « société de taille moyenne » et « petite société ». La variable « taille moyenne » est égale à 1 si la société est de taille moyenne, 0 sinon et la variable « petite société » est égale à 1 si la société est de petite taille, 0 sinon. Estimer une équation de la régression qui pourrait être utilisée pour prévoir le salaire annuel moyen des salariés étant donné le salaire annuel des employés à l'heure et la taille de l'entreprise.

d) Dans le cadre de l'équation estimée de la régression développée à la question (c), utiliser le test de Student pour déterminer si les variables indépendantes sont significatives au seuil de 0,05.

e) En vous basant sur vos résultats à la question (d), développer une équation estimée de la régression qui pourrait être utilisée pour prévoir le salaire annuel moyen des employés salariés à temps complet étant donné le salaire annuel moyen des employés rémunérés à l'heure et la taille de l'entreprise.

48. L'association nationale de basket (NBA) enregistre diverses statistiques sur chaque équipe. Six de ses statistiques sont le pourcentage de parties gagnées (% gagnées), le pourcentage de paniers marqués (% paniers), le pourcentage de tirs à trois points réussis (% 3pts), le pourcentage de lancers francs réussis (% lancers), le nombre moyen de rebonds offensifs par jeu (RebondOff) et le nombre moyen de rebonds défensifs par jeu (RebondDéf). Les données contenues dans le fichier en ligne NBAStats fournissent les valeurs de ses statistiques pour les 30 équipes de la NBA au cours de la saison 2011-2012 (site Internet de ESPN, 3 octobre 2012). Une partie des données est présentée ci-dessous.

Équipe	% gagnées	% paniers	% 3pts	% lancers	RebondOff	RebondDéf
Atlanta	60,6	45,4	37,0	74,0	9,9	31,3
Boston	59,1	46,0	36,7	77,8	7,7	31,1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Toronto	34,8	44,0	34,0	77,0	10,6	31,4
Utah	54,5	45,6	32,3	75,4	13,0	31,1
Washington	30,3	44,1	32,0	72,7	11,7	29,9

a) Développer une équation estimée de la régression qui peut être utilisée pour prévoir le pourcentage de parties gagnées étant donné le pourcentage de paniers marqués. Au seuil de 0,05, tester l'existence d'une relation significative.

b) Interpréter la pente de l'équation estimée de la régression développée à la question (a).

c) Développer une équation estimée de la régression qui peut être utilisée pour prévoir le pourcentage de parties gagnées étant donnés le pourcentage de paniers marqués, le pourcentage de tirs à 3 points réussis, le pourcentage de lancers francs réussis, le nombre moyen de rebonds offensifs par jeu et le nombre moyen de rebonds défensifs par jeu.

d) Supprimer toute variable indépendante qui ne serait pas significative au seuil de 0,05 de l'équation estimée de la régression développée en (c) et ré-estimer l'équation de la régression en ne conservant que les variables indépendantes significatives.

e) En supposant que l'équation estimée de la régression développée à la question (d) peut être utilisée pour la saison 2012-2013, prévoir le pourcentage de parties gagnées par une équipe dont les statistiques de jeu sont les suivantes : % paniers = 45 ; % 3pts = 35 ; RebondOff = 12 et RebondDéf = 30.

PROBLÈME 1 *La société Consumer Research*

La société Consumer Research est une agence indépendante qui effectue des recherches sur les attitudes des consommateurs et les comportements des firmes. Lors d'une étude, un client souhaitait connaître les caractéristiques des consommateurs permettant de prévoir le montant annuel des charges liées à la détention d'une carte de crédit. Des données sur le revenu annuel, la taille du ménage et le montant annuel des charges liées à la carte de crédit d'un échantillon de 50 consommateurs, ont été collectées. Ces données figurent dans le fichier en ligne intitulé Consumer.

Revenu (milliers de dollars)	Taille du ménage	Charge annuelle (en dollars)	Revenu (milliers de dollars)	Taille du ménage	Charge annuelle (en dollars)
54	3	4 016	54	6	5 573
30	2	3 159	30	1	2 583
32	4	5 100	48	2	3 866
50	5	4 742	34	5	3 586
31	2	1 864	67	4	5 037
55	2	4 070	50	2	3 605
37	1	2 731	67	5	5 345
40	2	3 348	55	6	5 370
66	4	4 764	52	2	3 890
51	3	4 110	62	3	4 705
25	3	4 208	64	2	4 157
48	4	4 219	22	3	3 579

Revenu (milliers de dollars)	Taille du ménage	Charge annuelle (en dollars)	Revenu (milliers de dollars)	Taille du ménage	Charge annuelle (en dollars)
27	1	2 477	29	4	3 890
33	2	2 514	39	2	2 972
65	3	4 214	35	1	3 121
63	4	4 965	39	4	4 183
42	6	4 412	54	3	3 730
21	2	2 448	23	6	4 127
44	1	2 995	27	2	2 921
37	5	4 171	26	7	4 603
62	6	5 678	61	2	4 273
21	3	3 623	30	2	3 067
55	7	5 301	22	4	3 074
42	2	3 020	46	5	4 820
41	7	4 828	66	4	5 149

Rapport

1. Utiliser les méthodes de statistiques descriptives pour résumer les données. Commenter les résultats.
2. Développer les équations estimées des régressions, en considérant tout d'abord le revenu annuel comme variable indépendante, puis la taille du ménage. Quelle variable est le meilleur facteur explicatif du montant annuel des charges liées à la carte de crédit ? Discuter vos résultats.
3. Développer une équation estimée de la régression avec, pour variables indépendantes, le revenu annuel et la taille du ménage. Discuter vos résultats.
4. Quel est le montant annuel des charges liées à la carte de crédit d'un ménage composé de trois personnes, disposant d'un revenu annuel de 40 000 dollars ?
5. Discuter de l'utilité d'ajouter d'autres variables indépendantes au modèle. Quelles variables supplémentaires pourraient être utiles ?

PROBLÈME 2 *Prévoir les gains des conducteurs de NASCAR*

Matt Kenseth a gagné la course Daytona 500 en 2012, la plus importante cours de la saison NASCAR. Sa victoire ne fut pas une surprise puisqu'il avait fini 4e lors de la saison 2011 avec 2 330 points, derrière Tony Stewart (2 403 points), Carl Edwards (2 403 points) et Kevin Harviwk (2 345 points). En 2011, il a gagné 6 183 580 dollars en gagnant trois pole positions (le pilote le plus rapide lors des qualifications), trois courses, en finissant dans les cinq premiers 12 fois et dans les dix premiers 20 fois. Le système de points de NASCAR en 2011 attribuait 43 points au vainqueur, 42 points au second, et ainsi de suite

Tableau 13.6 ***Résultats NASCAR pour la saison 2011***

Pilote	Points	Pole position	Victoires	Top 5	Top 10	Gains ($)
Tony Stewart	2403	1	5	9	19	6 529 870
Carl Edwards	2403	3	1	19	26	8 485 990
Kevin Harvick	2345	0	4	9	19	6 197 140
Matt Kenseth	2330	3	3	12	20	6 183 580
Brad Keselowski	2319	1	3	10	14	5 087 740
Jimmie Johnson	2304	0	2	14	21	6 296 360
Dale Earnhardt Jr.	2290	1	0	4	12	4 163 690
Jeff Gordon	2287	1	3	13	18	5 912 830
Denny Hamlin	2284	0	1	5	14	5 401 190
Ryan Newman	2284	3	1	9	17	5 303 020
Kurt Busch	2262	3	2	8	16	5 936 470
Kyle Busch	2246	1	4	14	18	6 161 020
Clint Bowyer	1047	0	1	4	16	5 633 950
Kasey Kahne	1041	2	1	8	15	4 775 160
A. J. Allmendinger	1013	0	0	1	10	4 825 560
Greg Biffle	997	3	0	3	10	4 318 050
Paul Menard	947	0	1	4	8	3 853 690
Martin Truex Jr.	937	1	0	3	12	3 955 560
Marcos Ambrose	936	0	1	5	12	4 750 390
Jeff Burton	935	0	0	2	5	3 807 780
Juan Montoya	932	2	0	2	8	5 020 780
Mark Martin	930	2	0	2	10	3 830 910
David Ragan	906	2	1	4	8	4 203 660
Joey Logano	902	2	0	4	6	3 856 010
Brian Vickers	846	0	0	3	7	4 301 880
Regan Smith	820	0	1	2	5	4 579 860
Jamie McMurray	795	1	0	2	4	4 794 770
David Reutimann	757	1	0	1	3	4 374 770
Bobby Labonte	670	0	0	1	2	4 505 650
David Gilliland	572	0	0	1	2	3 878 390
Casey Mears	541	0	0	0	0	2 838 320
Dave Blaney	508	0	0	1	1	3 229 210
Andy Lally	398	0	0	0	0	2 868 220
Robby Gordon	268	0	0	0	0	2 271 890
J. J. Yeley	192	0	0	0	0	2 559 500

jusqu'à un point au pilote qui finissait en 43e position. De plus, tout pilote qui avait un tour d'avance sur ses concurrents recevait un point de bonus, le pilote qui faisait le plus de tours recevait également un point de bonus supplémentaire et le vainqueur de la course

bénéficiait de trois points de bonus. Mais le maximum de points qu'un pilote pouvait gagner sur une course était de 48. Le tableau 13.7 fournit les données des 35 premiers pilotes sur la saison 2011 (site Internet de NASCAR, 28 février 2012).

Rapport

1. Supposez que vous vouliez prévoir les gains ($) en utilisant uniquement soit le nombre de pole positions gagnées, soit le nombre de victoires, soit le nombre de fois où le pilote est arrivé dans les 5 premiers, soit le nombre de fois où le pilote est arrivé dans les 10 premiers. Laquelle de ces quatre variables fournit le meilleur estimateur des gains ?
2. Développer une équation estimée de la régression qui peut être utilisée pour prévoir les gains ($) étant donnés le nombre de pole positions, le nombre de victoires, le nombre d'arrivées dans le top 5 et le nombre d'arrivées dans le top 10. Tester la significativité individuelle des variables explicatives et discuter de vos résultats et conclusions.
3. Créer deux nouvelles variables indépendantes ; Top 2-5 et Top 6-10. La première correspond au nombre de fois où le pilote a fini entre la seconde et la cinquième place et le seconde correspond au nombre de fois où le pilote a fini entre la sixième et la dixième place. Développer une équation estimée de la régression qui peut être utilisée pour prévoir les gains en utilisant les variables Pole positions, Victoires, Top 2-5 et Top 6-10. Tester la significativité individuelle des variables et discuter de vos résultats et conclusions.
4. Sur la base de vos résultats, quelle équation de régression recommanderiez-vous pour prévoir les gains ? Interpréter les coefficients estimés de cette équation.

PROBLÈME 3 *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, le 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 trois types de voitures (13 petites berlines, 20 berlines familiales et 21 berlines haut de gamme), incluant le prix (en dollars) de chaque voiture testée, sont fournies dans le fichier en ligne CarValues (site Internet de *Consumer Reports*, 18 avril 2012). Pour tenir compte de l'effet de la taille de la voiture, une variable qualitative à trois valeurs (petite berline, berline familiale et berline haut de gamme), utilisez les variables muettes suivantes : « Familiale » = 1 si la voiture est une berline familiale, 0 sinon et « Haut de gamme » = 1 si la voiture est une berline haut de gamme, 0 sinon.

Rapport

1. Considérez le coût au kilomètre comme la variable dépendante et développez une équation estimée de la régression avec les variables muettes Familiale et Haut de gamme comme variables indépendantes. Discutez de vos résultats.
2. Considérez le score de valeur comme variable dépendante et développez une équation estimée de la régression en utilisant le coût au kilomètre, la note attribuée lors des tests sur route, l'évaluation de la fiabilité du véhicule et les variables muettes Familiale et Haut de gamme comme variables indépendantes.
3. Supprimez toutes variables indépendantes non significatives dans l'équation estimée de la régression développée à la question (2) au seuil de 0,05. Après avoir supprimé ces variables, ré-estimer l'équation de la régression.
4. Supposez que quelqu'un déclare « les petites voitures sont une meilleure affaire que les voitures plus grandes. » Considérez que les données relatives aux petites berlines correspondent aux voitures les plus petites et que les voitures haut de gamme représentent les voitures les plus grandes. Votre analyse soutient-elle cette position ?

ANNEXE 13.1 RÉGRESSION MULTIPLE AVEC MINITAB

Dans la section 13.2, nous avons présenté l'output obtenu grâce à Minitab dans le cadre de la société de transport Butler. Dans cette annexe, nous décrivons les étapes nécessaires pour obtenir cet output. Premièrement, les données (cf. fichier en ligne Butler) doivent être enregistrées dans une feuille de calcul de Minitab. Les kilomètres parcourus sont enregistrés dans la colonne C1, le nombre de livraisons est enregistré dans la colonne C2 et la durée de trajet (en heures) dans la colonne C3. L'intitulé des colonnes correspond aux noms des variables « Miles », « Deliv » et « Time ». Dans les étapes suivantes, nous

faisons référence aux données en utilisant leur nom. Les étapes suivantes décrivent comment utiliser Minitab pour produire les résultats présentés à la figure 13.4.

Étape 1. Sélectionner le menu **Stat**
Étape 2. Sélectionner le menu **Regression**
Étape 3. Choisir **Regression**
Étape 4. Lorsque la boîte de dialogue **Regression** apparaît :
Entrer Time dans la boîte **Response**
Entrer Miles et Deliv dans la boîte **Predictors**
Cliquer sur **OK**

ANNEXE 13.2 RÉGRESSION MULTIPLE AVEC EXCEL

Dans la section 13.2, nous avons présenté l'output obtenu grâce à Minitab dans le cadre de la société de transport Butler. Dans cette annexe, nous décrivons les étapes nécessaires pour obtenir cet output avec les outils de régression d'Excel. Référez-vous à la figure 13.10 pour suivre la procédure. Premièrement, les intitulés des variables Numéro, Miles, Livraisons et Durée sont enregistrés dans les cellules A1:D1 d'une feuille de calcul et les données d'échantillon (cf. fichier en ligne Butler) dans les cellules B2:D11. Les numéros de 1 à 10 inscrits dans les cellules A2:A11 identifient chaque observation.

Les étapes suivantes décrivent comment utiliser les outils de la régression Excel dans le cadre de l'analyse d'une régression multiple.

Étape 1. Cliquer sur **Data** dans la barre des tâches
Étape 2. Dans le groupe **Analysis**, cliquer sur **Data Analysis**
Étape 3. Choisir **Regression** dans la liste des outils d'analyse
Étape 4. Lorsque la boîte de dialogue Regression apparaît :
Entrer D1:D11 dans la boîte **Input Y Range**
Entrer B1:C11 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** (pour identifier le coin gauche supérieur de la partie de la feuille de calcul qui contiendra l'output)
Cliquer sur **OK**

Dans l'output Excel présenté à la figure 13.10, le nom de la variable indépendante x_1 est Miles (cf. cellule A30) et le nom de la variable indépendante x_2 est Livraisons (cf. cellule A31). L'équation estimée de la régression est

$$\hat{y} = -0{,}8687 + 0{,}0611x_1 + 0{,}9234x_2$$

Notez que les outils de régression Excel dans le cadre d'une régression multiple sont quasiment identiques à ceux utilisés dans le cadre d'une régression linéaire simple.

La principale différence réside dans le fait qu'un plus large champ de cellules est nécessaire pour identifier les variables indépendantes.

	A	B	C	D	E	F	G	H	I	J
1	**Numéro**	**Miles**	**Livraisons**	**Durée**						
2	1	100	4	9,3						
3	2	50	3	4,8						
4	3	100	4	8,9						
5	4	100	2	6,5						
6	5	50	2	4,2						
7	6	80	2	6,2						
8	7	75	3	7,4						
9	8	65	4	6,0						
10	9	90	3	7,6						
11	10	90	2	6,1						
12										
13	RÉSUMÉ									
14										
15	*Statistiques de la régression*									
16	Mutiple R	0,9507								
17	R Square	0,9038								
18	Adjusted R Square	0,8763								
19	Standard Error	0,5731								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	2	21,6006	10,8003	32,8784	0,0003				
25	Residual	7	2,2994	0,3285						
26	Total	9	23,9							
27										
28		*Coefficients*	*Erreur type*	*Statistique t*	*Valeur p*	*Inférieur 95 %*	*Supérieur 95 %*	*Inférieur 99 %*	*Supérieur 99 %*	
29	Constante	−0,8687	0,9515	−0,9129	0,3916	−3,1188	1,3813	−4,1986	2,4612	
30	Miles	0,0611	0,0099	6,1824	0,0005	0,0378	0,0845	0,0265	0,0957	
31	Livraisons	0,9234	0,2211	4,1763	0,0042	0,4006	1,4463	0,1496	1,6972	
32										

Figure 13.9 ***Output Excel obtenu dans le cadre de l'exemple de la société Butler avec deux variables indépendantes.***

ANNEXE 13.3 RÉGRESSION MULTIPLE AVEC STATTOOLS

Dans cette annexe, nous montrons comment utiliser StatTools pour effectuer les calculs de l'analyse de la régression dans le cadre du problème de la société de transport Butler. Commencer par utiliser Data Set Manager pour créer un ensemble de données StatTools pour ces données en suivant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes décrivent comment utiliser StatTools pour obtenir les résultats de la régression.

Étape 1. Cliquer sur le bouton **StatTools** dans la 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 StatTools-Regression 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 **Miles**
Dans la colonne intitulée **I** sélectionner **Deliveries**
Dans la colonne intitulée **D** sélectionner **Time**
Cliquer sur **OK**

L'analyse de la régression apparaît alors.

La boîte de dialogue StatTools-Regression contient plusieurs options avancées pour effectuer des estimations par intervalle de prévision et produire des graphiques des résidus. L'aide de StatTools fournit les indications appropriées pour utiliser ces options.