Chapitre 8

Régression

Au Chapitre 1, nous avons présenté la droite des moindres carrés et le coefficient de corrélation comme techniques purement descriptives: on observe deux variables, *Y*, la variable dite *endogène*; et *X*, la variable *exogène*, qui doit servir à expliquer et prédire *Y*. Nous nous sommes contentés de décrire la relation entre deux variables *dans l'échantillon* et à proposer une équation destinée à exprimer la relation entre *X* et *Y*. Or l'échantillon n'a d'intérêt que dans la mesure où il reflète une réalité qui dépasse les données de l'échantillon. Ces données sont issues d'une population (par exemple, des maisons récemment vendues on note le prix *y* et la surface du plancher *x*); ou d'une expérience (quand, par exemple, on mesure la pression *y* d'une masse de gaz à différentes températures *x*). C'est la population dont elles sont issues ou le phénomène qui a généré les données qui importe. L'échantillon est une image de la population, mais une image pas nécessairement très fidèle. Le passage de l'échantillon à la population consiste à distinguer ce qui est le reflet de la population de ce qui n'est qu'un accident du hasard. Pour cela nous allons définir un *modèle* — une série de suppositions concernant le mécanisme par lequel les données ont été générées. Le modèle présenté ici est appelé *modèle de régression linéaire simple*.

8.1 Le modèle de régression linéaire simple

Nous illustrons le modèle à l'aide des données du tableau 8.1, qui présente, pour un ensemble de 101 maisons vendues dans la région de Montréal, les valeurs de deux variables :

x : La surface du plancher, en mètre carrés, est la variable *exogène* ;
y : Le prix à la vente, en milliers de dollars, la variable *endogène*.

La variable exogène (*x*) est celle qui sert à prédire la valeur de la variable endogène (*y*). Ces données sont représentées comme 101 points dans $\mathbb{R}^2$ dans la figure 8.1, à laquelle est également tracée la droite des moindres carrés.

Traitement descriptif

Rappelons que la droite des moindres carrés est la droite $y = b_0 + b_1 x$ qui minimise

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.1.1}$$

où $\hat{y}_i = b_0 + b_1 x_i$. Elle est donnée par

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{et} \quad b_0 = \bar{y} - b_1 \bar{x}, \tag{8.1.2}$$

où

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \text{et} \quad S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{8.1.3}$$

La force de la dépendance linéaire entre deux variables quantitatives est mesurée par le *coefficient de corrélation r*, défini par

$$r = \frac{S_{xy}}{S_x S_y} \tag{8.1.4}$$

Tableau 8.1

Prix de vente (y) *et surface du plancher* (x) *de* 101 *maisons vendues dans la région de Montréal*

x	y	x	y	x	y	x	y	x	y	x	y
200	319	1155	157	1300	299	1511	249	1900	312	2358	319
551	69	1191	199	1316	170	1520	314	1925	246	2452	639
551	69	1200	349	1344	130	1523	269	2000	699	2475	399
672	90	1209	175	1344	119	1528	299	2000	379	2500	321
760	178	1222	80	1364	205	1548	289	2000	226	2500	439
775	60	1225	142	1390	499	1600	279	2026	339	2500	479
775	89	1225	142	1400	140	1600	294	2100	369	2600	429
775	89	1225	153	1400	142	1600	379	2100	275	2925	495
800	97	1225	158	1400	180	1600	269	2100	329	3200	549
800	89	1240	219	1413	85	1600	239	2130	435	3318	450
800	89	1250	169	1413	85	1600	87	2157	700	3331	190
910	98	1250	184	1430	269	1620	201	2173	359	3376	339
960	89	1280	145	1450	249	1659	339	2188	329	3700	539
1050	86	1288	259	1485	239	1700	329	2200	499	3827	469
1056	229	1300	195	1500	200	1711	365	2200	499	3850	775
1100	184	1300	200	1500	249	1800	142	2300	385	5625	1050
1120	139	1300	289	1500	349	1864	310	2323	540		

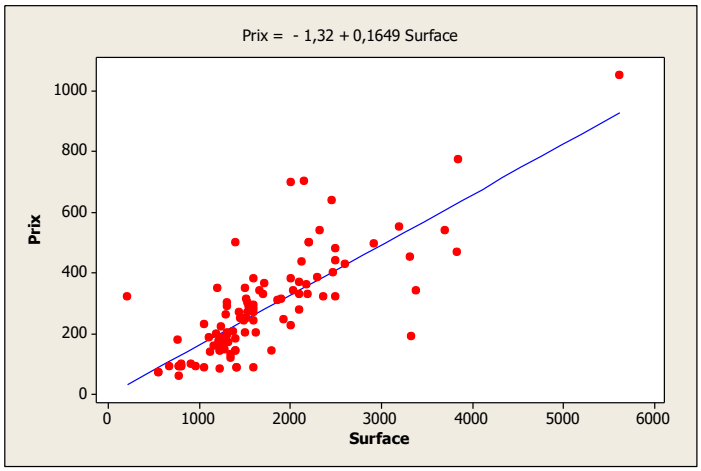
Exemple 8.1.1 *Calcul de la droite des moindres carrés*

Avec les données du tableau 8.1 déterminer la droite de régression est le coefficient de corrélation.

Solution Nous avons : $\bar{x} = 1709{,}07$, $\bar{y} = 280{,}6040$; $S_{xy} = 111\,632{,}2$; $S_x^2 = 676\,785$; $S_y^2 = 29\,740{,}78$; $b_o = -1{,}32$; $b_1 = 0{,}1649$. Le coefficient de corrélation est $r = 0{,}79$. ■

Figure 8.1

Relation entre le prix d'une maison (y) *et la surface du plancher* (x)

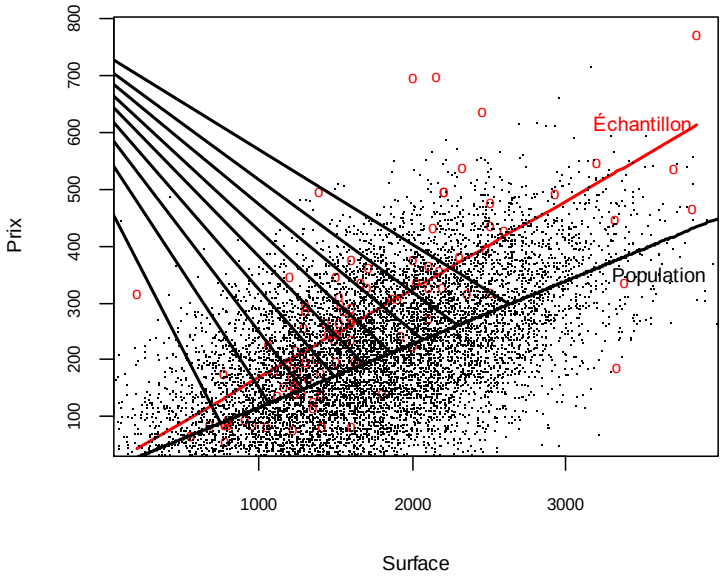


Le modèle

Le nuage de points que présente la figure 8.1 est un échantillon d'une population qui pourrait elle aussi être représentée par un (très grand) nuage de points correspondant à toutes les ventes susceptibles d'être réalisées dans la ville échantillonnée. La figure 8.2 présente l'échantillon observé ainsi que l'allure que pourrait prendre la population: un grand nuage révélant une dépendance qui peut elle aussi être représentée par une droite. On désignera cette droite (la «vraie» droite) par $y = \beta_o + \beta_1 x$. La droite des moindres carrés de l'échantillon $y = b_o + b_1 x$ déterminée à partir de l'échantillon serait alors une estimation de la vraie droite.

Figure 8.2

Relation entre le prix d'une maison (*y*) *et la surface du plancher* (*x*)
Population et échantillon



Afin de déterminer les propriétés de cet estimateur, nous ferons un ensemble de suppositions qui constituent le *modèle de régression linéaire simple.*

Dans le modèle de régression linéaire simple, les valeurs $x_1; \ldots ; x_n$ de x sont considérées comme des *constantes*, alors que les valeurs $y_1; \ldots ; y_n$ de y sont n variables aléatoires indépendantes. Nous supposons qu'elles sont de loi normale, de même variance

$$\mathrm{Var}(y_i) = \sigma^2 \tag{8.1.5}$$

Une façon compacte d'écrire le modèle est la suivante

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \ , \ i = 1, \ldots , n; \quad \varepsilon_i \sim \mathcal{N}(0 ; \sigma^2) \tag{8.1.6}$$

Notation *Afin d'éviter l'encombrement notationnel, nous avons présenté* (8.1.5) *sous une forme épurée qui pourrait, à l'occasion, porter à confusion. Si plusieurs variables sont définies dans un même contexte il est nécessaire de distinguer leurs variances par un indice:* σ_x^2 , σ_y^2 *, etc. Dans le cas de* (8.1.5)*, même la notation* σ_y^2 *n'aurait pas tout dit, car il ne s'agit pas de la variance de toutes les valeurs de* y_i *observées. Il s'agit de la variance des valeurs de Y qui correspondent à une même valeur donnée x. L'équation* (8.1.5) *devrait donc s'écrire plutôt comme*

$$\mathrm{Var}(y_i \mid x_i) = \sigma_{y.x}^2 \tag{8.1.5}'$$

Les espérances des y_i varient sont fonction linéaire des x_i, et la linéarité de la dépendance s'exprime par

$$E(y_i) = \beta_o + \beta_1 x_i,\ i = 1, \ldots, n \tag{8.1.7}$$

Remarque *Voici une façon d'interpréter les suppositions du modèle dans le contexte de l'exemple. Pour chaque valeur x_i de x, disons x_i = 1000 m^2, considérons l'ensemble de toutes les maisons dont la surface est de* 1000 m^2. *Les prix des maisons dans cette sous-population sont distribués selon une loi normale. Le prix moyen de ces maisons dépend de x : il est égal à* $\beta_o + \beta_1(1000)$. *La variance* σ^2 *est la dispersion des prix de toutes les maisons de* 1000 m^2. *On suppose que* cette variance est la même *pour toute sous-population constituée des maisons de même surface. Ainsi donc, la variance des prix des maisons de* 120 m^2 *serait également* σ^2. *C'est l'hypothèse dite* d'homoscédasticité. *Il est rare qu'elle soit vérifiée exactement en pratique, et dans cet exemple en particulier elle ne l'est assurément pas. Mais on ne s'attend pas à des effets très graves si les différences de variances ne sont pas très importantes.*
On laisse entendre, en utilisant le terme «population» ci-dessus, qu'il s'agit d'un ensemble fini de maisons; celles, par exemple, qui ont effectivement été vendues au courant de l'année dernière. Mais en fait le terme a un sens plus large. La vente des maisons, et leur prix, sont des phénomènes aléatoires : telle maison vendue aurait pu ne pas se vendre; telle autre, non vendue, aurait pu se vendre; le prix y d'une maison vendue, sujet à tous les aléas de ce genre de transaction, aurait pu être tout autre; etc. Bref, le plus souvent, la population qui nous intéresse est l'ensemble de toutes les transactions qui auraient pu *avoir lieu, et non celles qui se sont effectivement réalisées. Elle est donc infinie.* ■

8.2 Estimation de β_o, de β_1 et de σ^2

Nous avons 3 paramètres à estimer : β_o, β_1, et σ^2. Les estimateurs de β_1 et β_o sont précisément les valeurs b_1 et b_o (équations (8.1.2)) qui définissent la droite des moindres carrés. Nous désignerons b_1 et b_o désormais par $\hat{\beta}_o$ et $\hat{\beta}_1$, afin de mettre en évidence le fait que ces quantités jouent maintenant le rôle d'estimateurs:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}\,. \tag{8.2.1}$$

La valeur de *Y* peut être estimée pour chaque valeur de *x* par

$$\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i \tag{8.2.2}$$

L'estimateur de σ^2 est

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n \left(y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)\right)^2}{n-2} \tag{8.2.3}$$

Remarque *On peut justifier l'estimateur* (8.2.3) *intuitivement. On sait que* σ^2 *est la variance des erreurs* $\varepsilon_i = y_i - (\beta_o + \beta_1 x_i)$, *lesquelles sont de moyenne nulle. Un estimateur de* σ^2 *aurait donc été la moyenne des* ε_i^2, $\sum_{i=1}^n \varepsilon_i^2 / n$, *si les* ε_i *étaient connus. À défaut des* ε_i, *on utilise les* $\hat{\varepsilon}_i = y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)$. *Le numérateur de* $\hat{\sigma}^2$ *est donc* $\sum_{i=1}^n \hat{\varepsilon}_i^2$. *Le dénominateur doit cependant changer car le nombre de degrés de liberté de la somme de carrés* $\sum_{i=1}^n \hat{\varepsilon}_i^2$ *est n-2 et non plus n. Ce qui donne l'estimateur* $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}$, *l'estimateur proposé.* ■

On verra plus bas que ces estimateurs sont sans biais, on déterminera la variance de $\hat{\beta}_o$ et $\hat{\beta}_1$, et on montrera comment déterminer des intervalles de confiance et des tests d'hypothèse pour ces paramètres. Mais avant d'entrer dans ces détails, voici un exemple qui illustre le genre d'information que fournit un logiciel statistique. La plupart des logiciels présentent les résultats de base d'une analyse de régression d'essentiellement la même manière.

8.3 Propriétés des estimateurs

Estimateurs de β_o et de β_1

Les estimateurs $\hat{\beta}_o$ et $\hat{\beta}_1$ sont sans biais

$$E(\hat{\beta}_1) = \beta_1 \text{ et } E(\hat{\beta}_o) = \beta_o. \tag{8.3.1}$$

Sous l'hypothèse que les y_i sont de loi normale, $\hat{\beta}_o$ et $\hat{\beta}_1$ sont de loi normale.

Leur variance est

$$\text{Var}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \; ; \; \text{Var}(\hat{\beta}_o) = \sigma^2_{\hat{\beta}_o} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right). \tag{8.3.2}$$

Remarque *La formule de la variance $\text{Var}(\hat{\beta}_1)$ permet d'identifier les facteurs qui contribuent à améliorer l'estimateur. D'abord, on voit que plus σ^2 est petit, mieux c'est. Une relation forte, donc, est souhaitable non seulement parce qu'elle mène à de bonnes prédictions de y, mais aussi parce qu'elle améliore l'estimation des paramètres.*

Un deuxième facteur influence la qualité de l'estimateur : la valeur de la dispersion $\sum_{i=1}^n (x_i - \bar{x})^2$: la précision est d'autant meilleure que la somme $\sum_{i=1}^n (x_i - \bar{x})^2$ est grande. Cette somme, que l'on peut écrire comme $(n-1)S_x^2$, croît avec la taille de l'échantillon, comme il se doit. Mais elle croît aussi avec la dispersion des x. Cela veut dire que, dans la mesure où l'on peut contrôler les valeurs de x, on a intérêt à les choisir aussi dispersées que possible. ■

Estimateur de σ^2

L'estimateur de σ^2 est aussi sans biais:

$$E(\hat{\sigma}^2) = \sigma^2 \tag{8.3.3}$$

Quant à sa distribution, nous avons, sous l'hypothèse que normalité, la propriété suivante:

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \tag{8.3.4}$$

Estimation des variances de $\hat{\beta}_1$ et de $\hat{\beta}_o$

Nous devons estimer les variances de $\hat{\beta}_o$ et $\hat{\beta}_1$ ce qui se fait, naturellement, en substituant $\hat{\sigma}^2$ à σ^2 aux expressions (8.3.2):

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \quad \text{et} \quad \hat{\sigma}^2_{\hat{\beta}_o} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right). \tag{8.3.5}$$

Ces estimateurs sont sans biais également. Cela découle du fait que $\hat{\sigma}^2_{\hat{\beta}_1}$ et $\hat{\sigma}^2_{\hat{\beta}_o}$ sont des fonctions linéaires de $\hat{\sigma}^2$, qui est lui-même sans biais.

8.4 Distribution des statistiques de tests

Les statistiques $Z_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ et $Z_o = \dfrac{\hat{\beta}_o - \beta_o}{\sigma_{\hat{\beta}_o}}$ sont de loi $\mathcal{N}(0\,;\,1)$. Mais lorsqu'on remplace les écarts-types au dénominateur par leur estimation, les variables qui en résultent suivent une loi de *Student* à n-2 degrés de liberté :

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2} \quad \text{et} \quad T_o = \frac{\hat{\beta}_o - \beta_o}{\hat{\sigma}_{\hat{\beta}_o}} \sim t_{n-2} \tag{8.4.1}$$

Tests d'hypothèses

Les tests d'hypothèses pour β_1 et β_o se font de la même façon que les tests pour une moyenne μ. Une région critique de taille α pour tester

$$H_o : \beta_1 = b$$

est donnée par

$$\left| \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}} \right| > t_{n-2\,;\,\alpha/2} \tag{8.4.2}$$

De même, une région critique de taille α pour tester

$$H_o : \beta_o = a$$

est donnée par

$$\left| \frac{\hat{\beta}_o - a}{\hat{\sigma}_{\hat{\beta}_o}} \right| > t_{n-2\,;\alpha/2} \tag{8.4.3}$$

Intervalles de confiance

Les intervalles de confiance à 100(1- α) % pour β_o et β_1 sont donnés par

$$\hat{\beta}_o - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_o} \le \beta_o \le \hat{\beta}_o + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_o} \tag{8.4.4}$$

et

$$\hat{\beta}_1 - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1} \tag{8.4.5}$$

où $t_{n-2\,;\,\alpha/2}$ est le point critique correspondant à une loi de *Student* à n-2 degrés de liberté.

Il est rare qu'on ait à tester des hypothèses concernant β_o ; mais on s'intéressera toujours à β_1, normalement pour tester l'hypothèse que $\beta_1 = 0$.

Exemple 8.4.1 *Test et intervalles de confiance pour les coefficients*

Avec les données du tableau 8.1,

a) tester l'hypothèse que $\beta_1 = 0$;
b) déterminer un intervalle de confiance pour β_1 ;
c) déterminer un intervalle de confiance pour β_o.

Solutions

a) On a $\hat{\sigma} = 106{,}9675$; $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 67\,678\,520$. Donc $\hat{\sigma}_{\hat{\beta}_1} = \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{101}(x_i - \bar{x})^2}} = 0{,}01300$.

La statistique de test est $T = \dfrac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = 12{,}69$, ce qui est certainement significatif (le point critique pour un test à 5 % est $t_{99\,;0{,}025} = 1{,}9842$.) Le seuil expérimental est à toute fin pratique nul.

b) Intervalle de confiance pour β_1 :

$$[\,\hat{\beta}_1 - t_{99;0{,}025}\hat{\sigma}_{\hat{\beta}_1}\ ;\ \hat{\beta}_1 + t_{99;0{,}025}\hat{\sigma}_{\hat{\beta}_1}\,] = [0{,}139\,1451\ ;\ 0{,}190\,7466].$$

c) $\hat{\sigma}_{\hat{\beta}_o} = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{101}(x_i - \bar{x})^2}} = 24{,}64149$.

Intervalle de confiance pour β_o : $[\,\hat{\beta}_o - t_{99;0{,}025}\hat{\sigma}_{\hat{\beta}_o}\,;\hat{\beta}_o + t_{99;0{,}025}\hat{\sigma}_{\hat{\beta}_o}\,] = [-50.22\ ;\ 47.57]$. ■

Remarque *Le test de l'hypothèse* $H_o : \beta_1 = 0$ *est fondamental et doit précéder toute autre analyse, car c'est l'hypothèse selon laquelle il n'y a pas vraiment de relation entre les variables. Ce que nous avons conclu sans aucun doute raisonnable, c'est que la relation observée dans l'échantillon reflète une relation réelle dans la population. Autre façon habituelle d'exprimer la conclusion :* « $\hat{\beta}_1$ *est significativement différent de* 0 (*et donc positif*)».

Le rejet de H_o *était prévisible étant donné le coefficient de corrélation relativement satisfaisant de* 0,73. *Mais si le degré de confiance reflète aussi la taille de l'échantillon, et s'il est si fort, c'est aussi parce que l'échantillon est grand.* ■

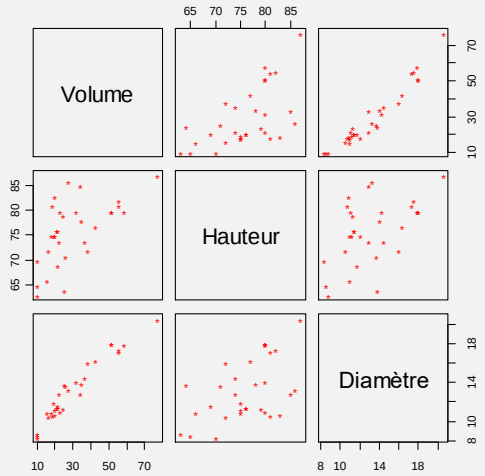
Les analyses de régression exigent d'importants calculs, d'autant plus que dans certaines situations plusieurs essais sont nécessaires afin de choisir le meilleur modèle. L'utilisation d'un logiciel est donc indispensable. Les sorties d'ordinateur présentées ici proviennent d'un logiciel spécialisé. Mais Excel peut également être utilisé. Quelques instructions sont présentées en annexe à cette fin.

Exemple 8.4.1 [Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) *The Minitab Student Handbook*. Duxbury Press.] Les données suivantes portent sur les dimensions d'un échantillon de 31 cerisiers, le but visé étant d'estimer le volume du tronc à partir de deux quantités facilement mesurables: la hauteur de l'arbre, et son diamètre. Le volume est exprimé en pieds cubes, la hauteur en pieds et le diamètre en pouces à une hauteur donnée du sol (4 pieds, 6 pouces).

Tableau 8.2
Diamètre, hauteur et volume de 31 troncs de cerisiers

Diamètre	Hauteur	Volume	Diamètre	Hauteur	Volume	Diamètre	Hauteur	Volume
8,3	70	10,3	11,3	79	24,2	14,0	78	34,5
8,6	65	10,3	11,4	76	21,0	14,2	80	31,7
8,8	63	10,2	11,4	76	21,4	14,5	74	36,3
10,5	72	16,4	11,7	69	21,3	16,0	72	38,3
10,7	81	18,8	12,0	75	19,1	16,3	77	42,6
10,8	83	19,7	12,9	74	22,2	17,3	81	55,4
11,0	66	15,6	12,9	85	33,8	17,5	82	55,7
11,0	75	18,2	13,3	86	27,4	17,9	80	58,3
11,1	80	22,6	13,7	71	25,7	18,0	80	51,5
11,2	75	19,9	13,8	64	24,9	18,0	80	51,0
						20,6	87	77,0

S'il fallait qu'on n'utilise qu'une seule des variables «Diamètre» et «Hauteur» pour prédire le volume, le graphique suivant suggère un choix :



Il est clair que le diamètre prédira mieux le volume que la hauteur. Pour montrer comment se manifeste, en chiffres, la supériorité de la variable «Diamètre», nous traiterons les deux choix.

Modèle 1 Considérons la hauteur comme variable exogène :

$$E(\text{Volume} \mid \text{hauteur}) = \beta_o + \beta_1(\text{hauteur})$$

Voici ce que fournit typiquement le logiciel:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835 **
hauteur       1.5433     0.3839   4.021 0.000378 ***
Residual standard error : 13.4 on 29 degrees of freedom
Multiple R-squared : 0.3579
```

La première colonne («`Estimate`») donne les estimations de β_o et de β_1: On a donc $\hat{\beta}_o = -87,1236$ *et* $\hat{\beta}_1 = 1,5433$ et pour une hauteur donnée, le volume est exprimé par volume $= -87,1236 + 1,5433 \times$hauteur.

La deuxième colonne («`Std. Error`») fournit les estimations des écarts-types de $\hat{\beta}_o$ et de $\hat{\beta}_1$: $\hat{\sigma}_{\hat{\beta}_o} = 29,2731$ et $\hat{\sigma}_{\hat{\beta}_1} = 0,3839$.

La troisième colonne («`t value`») normalise les estimations: $\hat{\beta}_o / \hat{\sigma}_{\hat{\beta}_o} =$ `-2,976` et $\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1} =$ `4,021`. Ces quantités sont les valeurs de variables de loi de *Student* sous les hypothèses $\beta_o = 0$ *et* $\beta_1 = 0$, respectivement. La première de ces hypothèses est rarement intéressante, mais l'hypothèse que $\beta_1 = 0$ est fondamentale. Car si $\beta_1 = 0$, la droite de régression est horizontale, l'espérance de y_i est égale à β_o, quelle que soit la valeur de x, et donc x n'est d'aucune utilité dans la prédiction de y.

Pour tester l'hypothèse que $\beta_1 = 0$ on peut comparer la valeur $\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1} =$ `4,021` à un point critique tiré de la table de la loi de *Student*. Dans le cas présent (voir la section 8.4), le point critique à 5 % est 2,045, ce qui signifie qu'on peut avec confiance rejeter l'hypothèse que $\beta_1 = 0$, puisque 4,021 > 2,045. Mais on peut faire mieux, puisque, la valeur p, dans la dernière colonne, est très faible — `0.000378` — et montre qu'on aurait pu rejeter l'hypothèse à un niveau bien inférieur à 5 %. Et conclure que la droite de régression n'est pas horizontale.

La ligne «`Residual standard error`» fournit l'estimation de σ *:* $\hat{\sigma}$ `=13,4`. Ceci représente la dispersion des volumes de tous les cerisiers de même hauteur (et non de tous les cerisiers de la population).

Finalement, «`Multiple R-squared : 0.3579`» est le carré du coefficient de corrélation, et donc r = 0,5982.

Modèle 2 Est-ce que le diamètre de l'arbre prédirait mieux le volume ? Considérons le modèle

$$E(\text{Volume} \mid \text{Diamètre}) = \beta_o + \beta_1(\text{Diamètre})$$

L'analyse suivante le confirme :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
diametre      5.0659     0.2474   20.48  < 2e-16 ***
Residual standard error : 4.252 on 29 degrees of freedom
Multiple R-squared : 0.9353
```

La valeur de r^2 est nettement supérieure et le niveau de signification est encore plus convaincant. On peut difficilement améliorer cette performance, mais il est toujours préférable d'étayer le processus de modélisation sur des connaissances préalables quand on en a. Dans le cas présent, nous pourrions considérer qu'un tronc est grossièrement cylindrique ou conique. Dans ce cas, le volume devrait être une fonction linéaire de la superficie d'une coupe multipliée par la hauteur. On pourrait donc considérer comme variable exogène

$$\text{d2h} = (\text{diamètre}^2)(\text{hauteur})$$

Modèle 3 Considérons donc le modèle

$$E(\text{Volume} \mid \text{d2h}) = \beta_o + \beta_1(\text{d2h})$$

On obtient les résultats suivants :

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.977e-01  9.636e-01  -0.309     0.76
d2h          2.124e-03  5.949e-05  35.711   <2e-16 ***
Residual standard error : 2.493 on 29 degrees of freedom
Multiple R-squared : 0.9778
```

On constate une certaine amélioration.

L'argument qui suggère le modèle 3 devrait normalement mener à un modèle sans constante β_o. La valeur p de 0,76 rend cette hypothèse plausible. On peut en fait l'imposer au modèle (voir l'exercice 8.12).

Modèle 3' Voici les résultats du modèle

$$E(\text{Volume} \mid \text{d2h}) = \beta_1(\text{d2h})$$

```
     Estimate Std. Error t value Pr(>|t|)
d2h 2.108e-03  2.722e-05   77.44   <2e-16 ***
Residual standard error : 2.455 on 30 degrees of freedom
```

Un ajustement satisfaisant. Remarquez que le test effectué dans le tableau est sans intérêt. Quand le modèle stipule que E(y) = $\beta_1 x$, l'hypothèse que $\beta_1 = 0$ entraîne que E(y) = 0, ce qui n'aurait pas de sens. On juge de la validité du modèle à la valeur de l'écart-type de 2,455 qui est à peine supérieure à celle (2,493) dans le modèle précédent.

Si on suppose un tronc conique, alors le rayon de l'arbre à la base doit être $r = uh/(h\text{-}4{,}5)$, où u est le rayon observé à la hauteur de 4,5 (en pieds), et h est la hauteur. Le volume devrait dont être égal à $\pi r^2 h/3$.

Modèle 4 Prenons donc $z = \pi r^2 h$ pour variable exogène. Il faudrait ici, si notre modèle est bon, que $\beta_o = 0$ et $\beta_1 = 1/3$. Le modèle devrait donc être E(Volume | z) = (1/3)z . On verra s'il est corroboré par les données. Considérons, d'abord, ce que donne une régression dans laquelle la supposition $\beta_o = 0$ n'est pas imposée, et dont le modèle est donc

$$E(\text{Volume} \mid z = \beta_o + \beta_1 z)$$

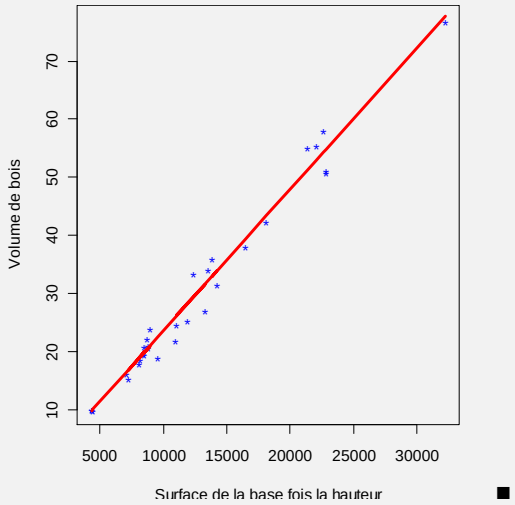
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.696197   0.984240  -0.707    0.485
z            0.350076   0.009912  35.318   <2e-16 ***
Residual standard error : 2.52 on 29 degrees of freedom
Multiple R-squared : 0.9773
```

Remarquez que le coefficient β_1, dont la valeur devrait être voisine de 1/3, selon le modèle conique, est estimé à $\hat{\beta}_1 = 0{,}35$, ce qui crédite le modèle.

Modèle 4' On ne rejette pas l'hypothèse que $\beta_o = 0$, et on pourrait à la rigueur l'éliminer du modèle, ce qui donnerait ceci :

```
    Estimate Std. Error t value Pr(>|t|)
z12  0.34385    0.00452   76.07   <2e-16 ***
Residual standard error: 2.499 on 30 degrees of freedom
```

Il n'est pas évident, cependant, qu'on veuille éliminer la constante β_o du modèle, puisqu'il n'est pas démontré que $\beta_o = 0$ et de toute façon, sa présence ne peut pas nuire. En fin de compte, le modèle 4 semble le plus approprié. Voici, finalement, une image de la relation entre le volume est z dans le modèle 4 :



■

8.5 Fonctions linéaires de β_o et de β_1 – intervalles de confiance

On peut estimer toute fonction linéaire $m = c_o\beta_o + c_1\beta_1$. Un estimateur sans biais est

$$\hat{m} = c_o\hat{\beta}_o + c_1\hat{\beta}_1,$$

et la variance de l'estimateur est

$$\mathrm{Var}(\hat{m}) = \sigma^2_{\hat{m}} = \sigma^2\left[\frac{c_o^2}{n} + \frac{(c_1 - c_o\bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{8.5.1}$$

Cette variance peut être estimée sans biais par

$$\hat{\sigma}^2_{\hat{m}} = \hat{\sigma}^2\left[\frac{c_o^2}{n} + \frac{(c_1 - c_o\bar{x})^2}{\sum(x_i - \bar{x})^2}\right], \tag{8.5.2}$$

et

$$\frac{\hat{m} - m}{\hat{\sigma}_{\hat{m}}} \sim t_{n\text{-}2} \tag{8.5.3}$$

Cas particulier La fonction $\mu_{y.x} = \beta_o + \beta_1 x$ représente la moyenne des y qui correspondent à une valeur donnée x, et $\hat{\mu}_{y.x} = \hat{\beta}_o + \hat{\beta}_1 x$ est une estimation de cette moyenne. La variance de cet estimateur est donnée par

$$\sigma^2_{\hat{\mu}_{y.x}} = \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{8.5.4}$$

Cette variance est estimée par

$$\hat{\sigma}^2_{\hat{\mu}_{y.x}} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{8.5.5}$$

On peut déterminer un test d'hypothèse pour μ_x comme on le fait pour une moyenne. L'estimateur $\hat{\mu}_{y.x}$ est de loi normale et la statistique $(\hat{\mu}_{y.x} - \mu_{y.x})/\hat{\sigma}_{\hat{\mu}_{y.x}}$ est de loi de *Student* :

$$\frac{\hat{\mu}_{y.x} - \mu_{y.x}}{\hat{\sigma}_{\hat{\mu}_{y.x}}} \sim t_{n-2} \tag{8.5.6}$$

Donc on rejette l'hypothèse.

$$H_o : \mu_{y.x} = \mu_{y.x_o}$$

si et seulement si

$$\left|\frac{\hat{\mu}_{y.x} - \mu_{y.xo}}{\hat{\sigma}_{\hat{\mu}_{y.x}}}\right| > t_{\alpha/2;n-2} \tag{8.5.7}$$

Un intervalle de confiance à 100(1-α)% est donné par

$$\hat{\mu}_{y.x} - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \leq \mu_{y.x} \leq \hat{\mu}_{y.x} + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \tag{8.5.8}$$

Limites de prédiction

Notez bien que l'intervalle ci-dessus est un intervalle de confiance pour la *moyenne* des *y* qui correspondent à une valeur donnée *x*. On peut affirmer, avec 100(1-α)% de confiance, que cette moyenne satisfait les inégalités suivantes :

$$\hat{\mu}_{y.x} - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \leq \mu_{y.x} \leq \hat{\mu}_{y.x} + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \qquad (8.5.9)$$

Mais on ne prétend pas que le *prochain y* qui correspond à *x* se situera entre ces deux bornes. Pour déterminer des bornes dans lesquelles une *valeur future* de *y* se trouvera avec une probabilité de 1 - α, nous raisonnons comme ceci. Si y_x est la future observation qui correspond à la valeur *x*, notre prédiction de y_x, que nous désignons par $\hat{y}_x$, sera identique à notre estimation $\hat{\mu}_{y.x}$ de la moyenne au point *x*. L'écart $y_x - \hat{y}_x$ satisfait

$$E(y_x - \hat{y}_x) = 0 ; \quad Var(y_x - \hat{y}_x) = Var(y_x) + Var(\hat{y}_x). \qquad (8.5.10)$$

La variance $Var(y_x) = \sigma^2$ est estimée par $\hat{\sigma}^2$ et $Var(\hat{y}_x) = Var(\hat{\mu}_x)$ est estimée par la formule donnée plus haut. Donc

$$\hat{\sigma}^2_{y_x - \hat{y}_x} = \hat{\sigma}^2 + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}\right) = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}\right) \qquad (8.5.11)$$

Les limites de prédiction à 100(1-α)% sont donc

$$\hat{y}_x - t_{n-2;\alpha/2}\hat{\sigma}_{y_x - \hat{y}_x} \leq y_x \leq \hat{y}_x + t_{n-2;\alpha/2}\hat{\sigma}_{y_x - \hat{y}_x} \qquad (8.5.12)$$

Là on peut affirmer avec 100(1-α)% de sécurité que la prochaine observation se situera entre les deux bornes.

Exemple 8.5.1 Avec les données du tableau 8.1 et à un niveau de confiance de 95 %, voici l'intervalle de confiance et les limites de prédiction suivantes pour une maison de 1750 m² :

Estimation des coefficients: $\hat{\beta}_o$ = -1,32432 ; $\hat{\beta}_1$ = 0,1649.

Estimation de $\mu_{1750} = \beta_o + \beta_1(1750)$: $\hat{\mu}_{1750} = \hat{\beta}_o + \hat{\beta}_1(1750)$ = 287,329 ;

Estimation de σ : $\hat{\sigma}$ = 106,9675 ;

Estimation de l'écart-type de l'estimateur : $\hat{\sigma}_{\hat{\mu}_{1750}}$ = 10,66 ;

Estimation de l'écart-type de $y_{1750} - \hat{y}_{1750}$: $\hat{\sigma}_{y_{1750} - \hat{y}_{1750}}$ = 107,5 ;

Point critique (α = 0,05) : $t_{99\ ;0,025}$ = 1,9842 ;
Intervalle de confiance : [266,184 ; 308,475] ;
Limites de prédiction : [74,032 ; 500,627].

Conclusion : On estime que le prix moyen des maisons de 1750 m² est de 287 329 $ et on affirme avec 95 % de confiance que cette moyenne se situe entre 266 184 $ et 308 475 $. Face à une maison de 1750 m², nous prédisons qu'elle se vendra à 287 329 $, mais tout ce que nous pouvons affirmer

avec 95 % de confiance, c'est que le prix auquel elle se vendra se situe entre 74 032 $ et 500 627 $!

■

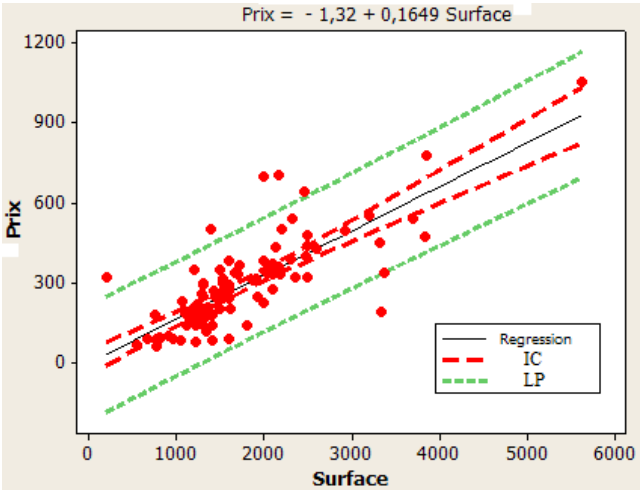
Remarque *La marge d'erreur* (*la demi-largeur de l'intervalle de confiance*) *dans l'estimation de* μ_{1750} *est de* 21 146 $. *En général, la marge d'erreur dépend de l'écart entre x et la moyenne* $\bar{x}$ *: plus l'écart est grand, plus la marge d'erreur est grande. Dans l'exemple nous avons fait un choix favorable, puisque* 1750 *n'est pas trop éloignée de la moyenne,* 1709. *La marge d'erreur dans l'estimation de* μ_{4000} *aurait été de* 62 761 $.

Ce qui est plus frappant dans l'exemple, c'est l'étonnante largeur de l'intervalle de prédiction. C'est un fait que, comme outil de prédiction, la régression ne fait pas de miracles, à moins que la relation soit très forte, et un coefficient de corrélation de 0,79 *n'est pas suffisant pour donner des prédictions très précises. Mais rappelons que le niveau de confiance exigé,* 95 %, *est très élevé. Si on s'était contenté d'un niveau de confiance de* 50 %, *les limites de prédiction auraient été* [214 556 ; 360 102]. *C'est mieux, mais encore trop large pour être utile.*

Avant de rejeter l'outil comme instrument de prédiction, il vaut mieux se rappeler qu'on serait encore plus démuni sans l'information que fournit la surface du plancher. Car si cette information n'était pas utilisée, et aucune autre non plus, la seule prédiction du prix que nous pourrions faire serait simplement la moyenne des prix. Les limites de prédiction seraient alors [246 559 ; 314 649] !

■

Les intervalles de confiance et les limites de prédiction peuvent être présentés pour tout *x* dans un graphique comme celui-ci :



Exemple 8.5.2 Tous les estimateurs définis ici peuvent être calculés par les logiciels. Voici un résumé de ce que peut produire le logiciel R (un logiciel statistique gratuit disponible sur le Web) pour les données du tableau 8.1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat{\beta}_0 = -1.32$	$\hat{\sigma}_{\hat{\beta}_0} = 24.6415$	$t_0 = \hat{\beta}_0 / \hat{\sigma}_{\hat{\beta}_0} = -.054$	$P(\|T_{n-2}\|>\|t_0\|) = 0.957$
surface	$\hat{\beta}_1 = 0.1649$	$\hat{\sigma}_{\hat{\beta}_1} = 0.0130$	$t_1 = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1} = 12.686$	$P(\|T_{n-2}\|>\|t_1\|) =$ <2e-16

Residual standard error : $\hat{\sigma}$ =107 on n-2 = 99 degrees of freedom

Multiple R-squared : $r^2 = 0.6191$

Intervalles de confiance pour μ_x. Voici des intervalles de confiance : pour μ_{500} ; μ_{2000} et μ_{3000}, *à* 95 % de confiance :

surfaces	fit	lwr	upr
500	81148	43474	118822
2000	328565	306153	350977
3000	493510	454076	532944

Prenons la deuxième ligne et précisons le sens de l'intervalle [306 153 ; 350 977]. Il veut dire ceci : nous affirmons, avec 95 % de confiance, que le prix moyen des maisons de 2000 m^2 se situe entre 306 153 $; 350 977 $. On constate que les intervalles de confiance pour μ_x sont particulièrement larges lorsque x est éloignée de la surface moyenne de 1709 m^2 (comme lorsque x = 500 ou 3000 m^2) et un peu moins larges lorsque x s'en approche (comme avec une surface de 1000 m^2). Dans tous les cas, cependant, l'intervalle est plutôt large, en partie parce que le niveau de confiance exigé (95 %) est plutôt ambitieux. Le tableau suivant montre l'effet d'une réduction du niveau de niveau de confiance à 80 % :

surfaces	fit	lwr	upr
500	81148	56652	105644
2000	328565	313993	343138
3000	493510	467870	519150

On détermine maintenant des intervalles de prédiction, à un niveau de confiance de 80 % :

surfaces	fit	lwr	upr
500	81148	-59014	221311
2000	328565	189793	467338
3000	493510	353143	633877

Interprétons maintenant le deuxième intervalle, [189 793 ; 467 338]. Il signifie que nous pouvons affirmer avec 80 % de confiance que le prix auquel se vendra une prochaine maison de 2000 m^2 se situe entre 189 793 $ et 467 338 $. La largeur de cet intervalle est déconcertante (et encore plus le premier intervalle, dont la borne inférieure est négative !). Ceci reflète une réalité incontournable : nos prédictions, si elles ne sont basées que sur la surface du plancher, risquent d'être gravement erronées, et ce malgré un coefficient de corrélation assez élevé (0,79). Le fait est que le prix d'une maison dépend bien plus que de sa taille. ■

Le sens de r^2

Le dernier exemple montre que les prédictions peuvent être très imprécises, même lorsque le coefficient de corrélation est assez élevé. Que signifie, alors, un coefficient de corrélation élevé s'il ne garantit pas de bonnes prédictions ? C'est le *carré* r^2 du coefficient de corrélation qui s'explique clairement. Il représente une *réduction relative de variance*. Dans l'exemple, nous avons calculé la variance des prix, $\hat{\sigma}^2 = (106{,}9675)^2$. Mais on pourrait aussi calculer une autre variance, $S_y^2 = (172{,}455)^2$ qui, elle aussi représente la dispersion des prix. Quelle est la différence ? S_y estime la dispersion des prix de toutes les maisons de la population, indépendamment de leur surface, alors que $\hat{\sigma}$ estime la dispersion σ des prix des maisons de *même surface*. σ est donc inférieur ou égal à l'écart-type des prix de toutes les maisons de la population. Le fait de connaître la surface, donc, signifie qu'on a affaire à un ensemble de prix moins dispersé. Le carré du coefficient de corrélation r^2 est essentiellement la réduction relative de la variance lorsqu'on se limite à des maisons de même surface, *à peu près* $\frac{S_y^2 - \hat{\sigma}^2}{S_y^2}$. Il peut donc être élevé, lorsque la réduction est relativement importante, sans que cela n'entraîne une variance faible. Les quantités précises sont définies dans le prochain paragraphe.

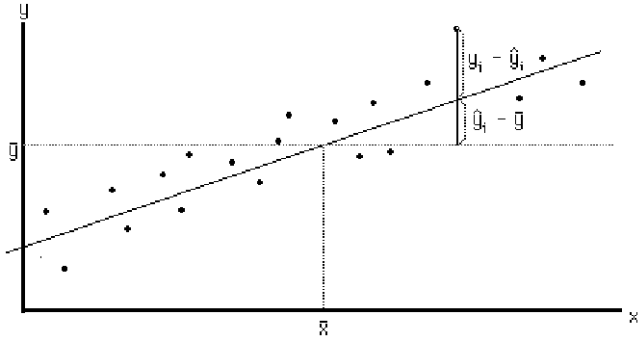
8.6 Analyse de variance

La somme des carrés $\sum_i (y_i - \bar{y})^2$, que nous appelons « somme des carrés *totale* » et désignons par SCT est une mesure de la dispersion totale des y, indépendamment des x. Cette somme de carrés

peut être décomposée en deux parties. La première, $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, appelée "somme des carrés *expliquée*" et notée SCE, est la partie de la dispersion des *y* qui est attribuable à la dispersion des *x*, donc «expliquée» par *x*. La deuxième, $\sum_{i=1}^{n}(y_i - \bar{y})^2$, appelée «somme des carrés résiduelle» et notée SCR, est la partie de la dispersion totale des *y* que l'on ne peut *pas* attribuer aux variations des *x*. Nous avons donc :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \qquad (8.6.1)$$

$$\text{SCT} \quad = \quad \text{SCE} \quad + \quad \text{SCR}$$



Graphiquement, SCE est la somme des carrés des distances verticales entre les points sur la droite des moindres carrés $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$ et les points sur la droite horizontale $y = \bar{y}$. Cette somme de carrés a tendance à être petite si la droite des moindres carrés s'approche d'une droite horizontale, c'est-à-dire, si les données ne témoignent pas d'une forte dépendance entre *y* et *x*. SCR est la somme des distances verticales entre les points du nuage et la droite des moindres carrés. Cette somme de carrés a tendance à être petite si les points sont rapprochés de la droite des moindres carrés, cas où la dépendance entre *y* et *x* est forte.

Remarques

1. SCR et $\hat{\sigma}^2$ *sont liés par la relation suivante* :

$$\hat{\sigma}^2 = \text{SCR}/(n\text{-}2)$$

Donc SCR *petit signifie que les* y_i *ont tendance à être* peu *dispersés par rapport à leur moyenne*

$$\beta_o + \beta_1 x_i,$$

ce qui se manifeste dans l'échantillon par un nuage de points rapproché de la droite des moindres carrés. Nous avons aussi la relation suivante entre $\hat{\beta}_1$ et SCR :

$$\text{SCR} = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

2. SCE *et* $\hat{\beta}_1$ *sont liés par la relation suivante* :

$$\text{SCE} = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Donc SCE *petit signifie que* $|\hat{\beta}_1|$ *est petit, et par conséquent que la droite est proche d'une droite horizontale. La décomposition est traditionnellement présentée sous la forme d'une table appelée «table d'analyse de variance», dans laquelle on indique aussi les «moyennes» de carrés, c'est-à-dire, les sommes de carrés divisées par le nombre de degrés de liberté.* ■

La table d'analyse de variance est également calculée par les logiciels. Le tableau contient les informations suivantes :

Table d'analyse de variance

Source	Degrés de liberté	Somme de carrés	Moyenne de carrés	Statistique F	Valeur p
Régression	1	SCE	MCE = SCE	f = MCE/MCR	P($F_{n-2;1} > f$)*
Erreur	n-2	SCR	MCR = SCR/(n-2)		
Total	n-1	SCT			

*$F_{n-2;1}$ désigne une variable de loi de Fisher à n-2 et 1 degrés de liberté

La colonne «Statistique F» ne sera pas discuté ici, mais le seuil expérimental est le seuil du test de l'hypothèse que $\beta_1 = 0$. Voici les valeurs de ce tableau produit le logiciel MINITAB pour les données traitées dans ce chapitre :

```
Analysis of Variance Table

Response: prix
           Df  Sum Sq Mean Sq F value    Pr(>F)
surface     1 1841316 1841316  160.93 < 2.2e-16 ***
Residuals  99 1132762   11442
```

Relation entre $\hat{\beta}_1$ *et le coefficient de corrélation*

Le coefficient de corrélation satisfait toujours

$$|r| \leq 1$$

et $|r| = 1$ si et seulement si il existe un nombre b tel que $y_i - \bar{y} = b(x_i - \bar{x})$ pour $i = 1, \ldots, n$, ce qui est équivalent à la condition $y_i = a + bx_i$ pour un certain a pour $i = 1, \ldots, n$. Donc les valeurs $r = 1$ et $r = -1$ dénotent une corrélation linéaire parfaite entre les x_i et les y_i. En comparant les expressions de $\hat{\beta}_1$ et de r, on constate que r et $\hat{\beta}_1$ sont de même signe et que $r = 0 \Leftrightarrow \hat{\beta}_1 = 0$. Nous avons la relation suivante :

$$r = \hat{\beta}_1 \frac{S_x}{S_y}, \text{ et donc } \hat{\beta}_1 = r \frac{S_y}{S_x} \qquad (8.6.2)$$

Donc $r > 0$ si et seulement si la droite des moindres carrés est de pente positive, et $r = 0$ si et seulement si la droite des moindres carrés est horizontale. Pour interpréter les valeurs intermédiaires de r, nous avons l'égalité suivante :

$$r^2 = \frac{\text{SCT-SCR}}{\text{SCT}} = \frac{\text{SCE}}{\text{SCT}} \qquad (8.6.3)$$

Ce qui signifie que r^2 est la *partie de la dispersion des y qui est expliquée par la dispersion des x*.

RÉSUMÉ

1 Le modèle de régression linéaire simple est $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0 ; \sigma^2)$, $i = 1, \ldots, n$.

2 Les estimateurs sont $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ et $\hat{\beta}_o = \bar{y} - \hat{\beta}_1\bar{x}$

3 $\hat{\beta}_1 \sim N(\beta_1 ; \sigma^2_{\hat{\beta}_1})$, où $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$; $\hat{\beta}_o \sim N(\beta_o ; \sigma^2_{\hat{\beta}_o})$, où $\sigma^2_{\hat{\beta}_o} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$.

4 L'estimateur de σ^2 est $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i-(\hat{\beta}_o+\hat{\beta}_1 x_i))^2}{n-2}$.

5 $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}$ et $\hat{\sigma}^2_{\hat{\beta}_o} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$

6 $T_1 = \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$ et $T_o = \frac{\hat{\beta}_o-\beta_o}{\hat{\sigma}_{\hat{\beta}_o}} \sim t_{n-2}$

7 Intervalles de confiance :

$$\hat{\beta}_o - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_o} \le \beta_o \le \hat{\beta}_o + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_o} \text{ et}$$

$$\hat{\beta}_1 - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1},$$

où $t_{n-2;\alpha/2}$ est le point critique correspondant à une loi de *Student* à n-2 degrés de liberté.

8 Un estimateur sans biais de $m = c_o\beta_o + c_1\beta_1$ est $\hat{m} = c_o\hat{\beta}_o + c_1\hat{\beta}_1$.

$$\text{Var}(\hat{m}) = \sigma^2_{\hat{m}} = \sigma^2\left[\frac{c_o^2}{n} + \frac{(c_1-c_o\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].$$

Cette variance peut être estimée sans biais par

$$\hat{\sigma}^2_{\hat{m}} = \hat{\sigma}^2\left[\frac{c_o^2}{n} + \frac{(c_1-c_o\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].$$

La statistique $\frac{\hat{m}-m}{\hat{\sigma}_{\hat{m}}}$ suit une loi t_{n-2}.

9 Un intervalle de confiance pour $\mu_{y.x} = \beta_o + \beta_1 x$ est donné par

$$\hat{\mu}_{y.x} - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \le \mu_{y.x} \le \hat{\mu}_{y.x} + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_{y.x}} \text{ où } \hat{\sigma}^2_{\hat{\mu}_x} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$$

10 Limites de prédiction : $\hat{y}_x - t_{n-2;\alpha/2}\hat{\sigma}_{y_x-\hat{y}_x} \le y_x \le \hat{y}_x + t_{n-2;\alpha/2}\hat{\sigma}_{y_x-\hat{y}_x}$, où

$$\hat{\sigma}^2_{y_x-\hat{y}_x} = \hat{\sigma}^2 + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

11 Décomposition de la somme des carrés totale (SCT) :

$$\sum_i (y_i-\bar{y})^2 = \sum_i (\hat{y}_i-\bar{y})^2 + \sum_i (y_i-\hat{y}_i)^2 \quad = \text{SCE} + \text{SCR}.$$

12 Table d'analyse de variance :

Source	Somme des carrés	d.l.	Moyenne des carrés
Régression	SCE = $\sum(\hat{y}_i - \bar{y})^2$	1	MCE = SCE/1
Résiduelle	SCR = $\sum(y_i - \hat{y}_i)^2$	n - 2	MCR = SCR/(n-2) = $\hat{\sigma}^2$
Total	SCT = $\sum(y_i - \bar{y})^2$	n - 1	MCT = SCT/(n-1) = S_y^2

Tableau 8.3

Pouls (*y*) *et Nombre d'années dans les plaines* (*x*)

Nombre d'années	Pouls	Nombre d'années	Pouls	Nombre d'années	Pouls	Nombre d'années	Pouls
1	88	13	68	18	80	25	72
6	64	10	72	11	76	26	68
5	68	15	88	11	60	10	60
1	52	18	60	21	64	19	74
1	72	2	60	24	64	18	72
19	72	12	72	14	68	10	56
5	64	15	84	25	76	1	64
25	80	16	64	32	60	43	72
6	76	17	72	5	76	40	92
13	60	10	64	12	88		

Tableau 8.4

Temps de finition d'une toile (T) et surface de la toile (S)

T	*S*	*T*	*S*	*T*	*S*	*T*	*S*
5,50	9,30	7,00	16,70	6,50	15,80	6,90	16,70
5,90	13,50	7,50	23,20	6,50	14,90	6,80	15,80
5,80	11,10	5,50	11,10	7,10	18,60	6,60	16,70
6,30	14,90	7,20	20,40	7,00	15,80		

Tableau 8.5

Habileté mathématique (H) et résultat à un examen d'algèbre (F)

F	*H*	*F*	*H*	*F*	*H*	*F*	*H*
36	9	32	18	59	26	79	33
23	10	44	20	58	28	74	34
22	13	52	22	72	30	78	36
36	15	51	23	87	31	99	38
49	16	83	24	86	32	85	40

Tableau 8.6

Poids de 45 *poussins aux âges de* 6, 10, *et* 21 *semaines*

#	Groupe	Sixième semaine	Dixième semaine	21e semaine	#	Groupe	Sixième semaine	Dixième semaine	21e semaine
1	1	64	93	205	24	2	73	114	233
2	1	72	103	215	25	2	74	106	309
3	1	67	99	202	26	2	72	98	150
4	1	67	87	157	27	3	73	102	256
5	1	60	106	223	28	3	82	129	305
6	1	74	124	157	29	3	77	111	147
7	1	71	112	305	30	3	85	134	341
8	1	68	96	98	31	3	87	158	373
9	1	63	81	124	32	3	76	116	220
10	1	84	139	175	33	3	68	83	178
11	1	62	88	205	34	3	74	109	290
12	1	60	67	96	35	3	78	109	272
13	1	79	128	266	36	3	79	120	321
14	1	72	89	142	37	4	85	124	204
15	1	62	71	157	38	4	84	126	281
16	1	58	73	117	39	4	96	157	200
17	2	86	163	331	40	4	78	117	196
18	2	77	95	167	41	4	82	120	238
19	2	73	103	175	42	4	79	123	205
20	2	74	68	74	43	4	80	125	322
21	2	78	124	265	44	4	85	128	237
22	2	74	114	251	45	4	84	122	264
23	2	73	100	192					