Chapitre 10

Régression multiple

Le modèle étudié dans ce chapitre— la régression multiple— fait appel à des notions théoriques qui dépassent le niveau de cet ouvrage; et son application exige des calculs qui peuvent difficilement se faire sans l'aide d'un logiciel. Donc la théorie ne sera développée que dans ses grandes lignes: les démonstrations de théorèmes, ainsi que certaines formules, seront omises. Et les formules de calcul ne seront pas toutes explicitées: pour celles que ne le seront pas, nous nous en remettons à Excel.

Ces compromis n'affecteront pas la compréhension. Nous continuerons à énoncer les théorèmes formellement, dans le langage mathématique familier des chapitres précédents: il sera encore question d'estimateurs, de lois d'échantillonnage, d'intervalles de confiance, et de tests d'hypothèses.

Pour quelques calculs relatifs à la régression multiple, Excel ne suffit pas. Il faudra alors recourir à un logiciel spécialisé. Il y en a plusieurs sur le marché et il importe peu lequel est choisi: le format de base des sorties de logiciel est à peu de choses près le même pour tous. Pour les lecteurs qui s'initient à l'emploi d'un logiciel statistique, nous recommandons **MegaStat**, un logiciel associé à Excel. Quelques indications sur l'utilisation de **MegaStat** sont présentées en annexe de ce chapitre.

10.1 Le modèle

La régression linéaire simple développée au chapitre 8 permet de prédire la valeur d'une variable endogène Y à partir d'une variable exogène x. Il est clair, cependant que la valeur d'une variable est généralement affectée par plusieurs facteurs. Considérons les maisons vendues dans une ville au cours d'une période donnée, disons un an, et qu'on s'intéresse au prix de vente. On peut prédire le prix d'une maison donnée à partir d'une variable exogène comme, par exemple, le nombre x de chambres à coucher. Mais le prix d'une maison, même pour un x fixe, peut varier beaucoup à cause des nombreux autres facteurs qui contribuent au prix: le quartier, l'âge, le nombre de salles de bain, etc. Une prédiction basée sur ces informations promet d'être plus précise qu'une prédiction basée sur le seul nombre de chambres à coucher.

Un modèle pour ce faire est une extension toute naturelle du modèle de régression simple, le modèle de *Régression multiple*. Supposons qu'on dispose de k variables exogènes, x_1, x_2,…, x_k, et une variable endogène Y. L'échantillon est constitué de n observations sur chacune des variables. On désigne par Y_i la i^e observation sur Y et x_{1i} ; x_{2i} ; … ; x_{ki} les i^e valeurs des variables x_1, x_2, …, x_k, respectivement. On désigne ces valeurs par le vecteur

$$\boldsymbol{x_i} = [x_{1i}\,;\, x_{2i}\,;\, \ldots\,;\, x_{ki}].$$

Le modèle stipule que, étant donné une valeur $\boldsymbol{x} = [x_1\,;\, x_2\,;\, \ldots\,;\, x_k]$ des variables exogènes, l'espérance de Y est une fonction linéaire de $\boldsymbol{x}$:

$$\mu_{y.\boldsymbol{x}} = \mathrm{E}(Y \mid \boldsymbol{x}) = \beta_o + \beta_1 x_1 + \ldots + \beta_k x_k,$$

alors que la variance est indépendante de $\boldsymbol{x}$:

$$\mathrm{Var}(Y \mid \boldsymbol{x}) = \sigma^2.$$

Plusieurs des hypothèses du modèle de régression simple sont retenues:

- Les n variables Y_i sont supposées indépendantes;
- La variance de Y_i pour une valeur donnée de $\boldsymbol{x}$ est indépendante de $\boldsymbol{x}$ (l'hypothèse d'homoscédasticité) ;
- Les valeurs x_{1i}, x_{2i}, …, x_{ki} sont des nombres fixes et non des variables aléatoires;
- Les intervalles de confiance et les tests d'hypothèses reposent sur la supposition que les Y_i sont de loi normale:

$$Y_i \sim \mathcal{N}(\mu_{y.\boldsymbol{x}}\,;\, \sigma^2).$$

10.2 Estimation des paramètres, intervalles de confiance et tests d'hypothèses

Estimateurs des paramètres

Nous avons maintenant $k + 2$ paramètres à estimer:

$$\text{les } k+1 \text{ composantes du vecteur } \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ et } \sigma^2.$$

Plusieurs approches différentes mènent aux mêmes estimateurs $\hat{\beta}_j$ des β_j. Pami elles se trouve le principe des moindres carrés (tout comme pour une régression simple):

Les estimateurs de $\beta_1, \beta_2, \ldots, \beta_k$ sont les valeurs $\hat{\beta}_o, \hat{\beta}_1, \cdots, \hat{\beta}_k$ qui minimisent la somme des carrés des erreurs

$$Q = \sum_{i=1}^{n} [y_i - (\beta_o + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})]^2 .$$

Les « formules » des $\hat{\beta}_j$ s'expriment en un langage matriciel que nous préférons éviter ici. Nous les omettons donc et laissons les logiciels effectuer le calcul.

L'estimation de la variance σ^2 est basée sur les écarts quadratiques $(y_i - \hat{y}_i)^2$, où

$$\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} .$$

Un estimateur sans biais de σ^2 est donné par

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (k+1)} .$$

Remarque L'estimation de la variance est ici conforme au principe utilisé jusqu'ici: $\sigma^2 = E[(Y_i - \mu_i)^2]$, où

$$\mu_i = E(Y_i \mid \boldsymbol{x}_i) = \beta_o + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} .$$

Il est donc naturel de l'estimer par une moyenne des carrés des écarts $y_i - \mu_i$. Cette moyenne n'étant pas connue, on la remplace par son estimation $\hat{y}_i$. Cependant, nous entendons par « moyenne » la somme $(y_i - \hat{y}_i)^2$ divisée non pas par le nombre de termes n mais plutôt par $n - (k+1)$. C'est le nombre de degrés de liberté de la distribution de $\hat{\sigma}^2$, et c'est ce qui fait de $\hat{\sigma}^2$ un estimateur sans biais.

Il existe un estimateur sans biais $\hat{\sigma}^2_{\hat{\beta}_j}$ de la variance $\sigma^2_{\hat{\beta}_j}$ de $\hat{\beta}_j$; nous n'en donnons pas la formule, mais nous énonçons ses propriétés.

Distribution des estimateurs

Sous l'hypothèse de normalité,

$$\hat{\beta}_j \sim N(\beta_j ; \sigma^2_{\hat{\beta}_j})$$

Donc $\hat{\beta}_j$ est sans biais.

Il existe une formule pour sa variance $\sigma^2_{\hat{\beta}_j}$, formule qui sera elle aussi omise.

Comme on le remarque souvent, l'estimateur de la variance $\hat{\sigma}^2$ suit, à une constante multiplicative près, une loi khi-deux:

Sous l'hypothèse que les y_i sont de loi normale,

$$\frac{[n - (k+1)]\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2_{n-(k+1)}$$

Remarque

Ce résultat explique pourquoi le diviseur de $\hat{\sigma}^2$ doit être $n - (k+1)$ pour que $\hat{\sigma}^2$ soit sans biais. L'espérance d'une variable de loi χ^2 est égale à son nombre de degrés de liberté.

$$\mathrm{E}\left[\frac{[n-(k+1)]\hat{\sigma}^2}{\sigma^2}\right] = n - k + 1 \Rightarrow \frac{n-(k+1)}{\sigma^2}\mathrm{E}(\hat{\sigma}^2) = n - k + 1 \Rightarrow \mathrm{E}(\hat{\sigma}^2) = \sigma^2 : \ \hat{\sigma}^2 \text{ est donc sans biais.}$$

Il existe aussi des estimateurs $\hat{\sigma}^2_{\hat{\beta}_j}$ sans biais des variances $\sigma^2_{\hat{\beta}_j}$.

Finalement, les statistiques à la base de la construction des intervalles de confiance et des tests d'hypothèses concernant les β_j suivent la loi de *Student*:

Sous l'hypothèse que les Y sont de loi normale,

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(k+1)}$$

Intervalles de confiance

Il s'ensuit du dernier résultat qu'un intervalle de confiance de niveau 100(1-α)% est donné par

$$\hat{\beta}_j - t_{n-k-1;\alpha/2}\hat{\sigma}_{\hat{\beta}_j} \le \beta_j \le \hat{\beta}_j + t_{n-k-1;\alpha/2}\hat{\sigma}_{\hat{\beta}_j}$$

Test d'hypothèse concernant β_j

Si β_{jo} est une valeur donnée, un test bilatéral de l'hypothèse

$$\mathrm{H_o} : \beta_j = \beta_{jo} \ vs \ \mathrm{H_1} : \beta_j \neq \beta_{jo}$$

est donné par:

$$\text{On rejette } \mathrm{H_o} \text{ si } \left|\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}\right| > t_{n-k-1;\alpha/2}$$

Les hypothèses que nous voudrons tester généralement sont

$$\mathrm{H_o} : \beta_j = 0 \ vs \ \mathrm{H_1} : \beta_j \neq 0.$$

La statistique de test est $t_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ et la région critique est

$$\text{On rejette } \mathrm{H_o} \text{ si } |t_j| > t_{n-k-1;\alpha/2}$$

Un critère analogue est le suivant:

On rejette $\mathrm{H_o}$ si

$$\mathrm{P}(|T_{n-(k+1)}| > t_j \mid \mathrm{H_o}) < \alpha.$$

où $T_{n-(k+1)}$ désigne une variable aléatoire de loi de *Student* à $n - (k+1)$ degrés de liberté.

La probabilité

$$p = \mathrm{P}(|T_{n-(k+1)}| > t_j \mid \mathrm{H_o})$$

est appelée *valeur p*.

Exemple 10.2.1 *Le prix d'une maison en fonction de son âge et du nombre de chambres à coucher*

Le tableau 10.1 présente des données sur la vente de $n = 43$ maisons. On se propose de développer une formule permettant de prédire le prix d'une maison à partir de son âge et du nombre de chambres à coucher qu'elle a. Les variables observées sont:

`Prix`: Le prix auquel la maison s'est vendue (en milliers de dollars)

`age`: L'âge de la maison
`cc`: Le nombre de chambres à coucher

Le modèle de régression linéaire est

$$\texttt{Prix} = \beta_o + \beta_1(\texttt{age}) + \beta_2(\texttt{cc})$$

Nous effectuons les calculs à l'aide de **MegaStat**. Le logiciel fournit un grand nombre de données, mais les principaux résultats sont contenus dans le tableau suivant:

Regression output					confidence interval	
variables	*coeffi-cients*	*std. error*	*t (df=40)*	*p-value*	*95% lower*	*95% upper*
Intercept	32.4148	10.8474	2.988	.0048	10.4914	54.3382
age	-0.4295	0.0870	-4.936	1.45E-05	-0.6053	-0.2536
cc	12.1198	3.4052	3.559	.0010	5.2376	19.0020

a) *Estimation des coefficients* β_o, β_1, et β_2.
La réponse découle de la colonne *coefficients*
Sont indiquées dans cette colonne les estimations des coefficients β_o, β_1, et β_2.
Ainsi donc: $\hat{\beta}_0 = 32,4147$; $\hat{\beta}_1 = -0,4295$; $\hat{\beta}_2 = 12,1198$. On estime le prix moyen des maisons d'âge x_1 et de x_2 chambres à coucher par $32,4147 - 0,4295x_1 + 12,1198x_2$. Pour une maison de 15 ans avec 2 chambres à coucher, cela donne $32,4147 - 0,4295(15) + 12,1198(2) = 50\ 212$ $

b) *Estimation des écarts-types* $\sigma_{\hat{\beta}_o}$, $\sigma_{\hat{\beta}_1}$ et $\sigma_{\hat{\beta}_2}$
La réponse découle de la colonne *std. error*
Cette colonne présente les estimations $\hat{\sigma}_{\hat{\beta}_j}$ des écarts-types $\sigma_{\hat{\beta}_j}$ des $\hat{\beta}_j$.
Donc $\hat{\sigma}_{\hat{\beta}_o} = 10,8474$; $\hat{\sigma}_{\hat{\beta}_1} = 0,0870$; $\hat{\sigma}_{\hat{\beta}_2} = 3,4052$.
Ces écarts-types serviront au calcul des statistiques de *Student* pour tester des hypothèses et déterminer des intervalles de confiance.

c) *Tests des hypothèses* $\beta_j = 0$.
La réponse découle de la colonne *t (df=40)*
Cette colonne présente les statistiques $t_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$. Sous l'hypothèse que $\beta_j = 0$, t_j suit une loi de *Student* à n-k-1= 40 degrés de liberté. Avec $\alpha = 0,05$, on rejette l'hypothèse si $|t_j| > t_{40;0,025} = 2,02$. On a $|t_1| = 0,0048 > 2,02$; $|t_2| = 4,936 > 2,02$; et $|t_3| = 3,559 > 2,02$. On rejette les 3 hypothèses, ce qui permet de conclure que $\beta_o > 0$, $\beta_1 < 0$, et $\beta_2 > 0$.
La colonne *p-value* permet d'éviter ces calculs, puisqu'elle donne les *p-valeurs* $P(|T_{40}| > |t_j|)$ sous chacune des hypothèses $\beta_j = 0$. On rejette l'hypothèse si $p < \alpha$. On constate qu'avec $\alpha = 0,05$ on peut rejeter les trois hypothèses. Bien plus encore: on sait qu'on aurait pu les rejeter avec un α beaucoup plus petit.

d) En termes concrets, comment s'expriment les conclusions en c)?
L'hypothèse que $\beta_o = 0$ est testée automatiquement par le logiciel, mais elle est peu pertinente. En général, on ne s'attend pas à ce qu'elle soit vraie.
Les hypothèses $\beta_1 = 0$ et $\beta_2 = 0$, par contre, sont fondamentales: ce sont des tests sur la pertinence des variables exogènes.
L'hypothèse que $\beta_1 = 0$ affirme que l'âge de la maison est sans effet sur le prix. Le test conclut le contraire: le prix de la maison décroît avec l'âge. L'inclusion de cette variable exogène améliore la prédiction du prix.
Même chose pour l'hypothèse que $\beta_2 = 0$. Conclure que $\beta_2 > 0$, c'est conclure que le prix de la maison

croît avec le nombre de chambres à coucher (ce qui coule de source).

e) *Intervalle de confiance à* 95 % *pour les* β_j.
Les intervalles de confiance sont présentés dans les colonnes `confidence interval`.

L'intervalle pour β_o est sans intérêt.

L'intervalle pour β_1: La valeur $\hat{\beta}_1$ = -0,4295 signifie que le prix de la maison décroît avec l'âge à un taux estimé de 429,5 $ par année. L'intervalle de confiance entoure cette estimation d'une marge d'erreur de sorte qu'on peut affirmer avec 95 % de confiance que le taux de décroissance se situe entre 253 $ et 605 $ (en arrondissant au dollar le plus proche).

L'intervalle pour β_2: La valeur estimée est $\hat{\beta}_2$ = 12,1198. On estime donc qu'une chambre à coucher de plus ajoute (en moyenne) 12 120 $ à la valeur d'une maison. Pour être sûr (à 95 %), on dira que cette « valeur ajoutée » et situe quelque part entre 5 238 $ et 19 000 $.

Intervalles de confiance pour $\mu_{y.x}$ *et limites de prédiction*

Pour une valeur donnée $\boldsymbol{x}_o = [x_{1o}, x_{2o}, \ldots, x_{ko}]$ des variables exogènes, l'espérance

$$\mu_{y.\boldsymbol{x}_o} = E(Y \mid \boldsymbol{x}_o) = \beta_o + \beta_1 x_{1o} + \ldots + \beta_k x_{ko}$$

est naturellement estimée par

$$\hat{\mu}_{y.\boldsymbol{x}_o} = \hat{\beta}_o + \hat{\beta}_1 x_{1o} + \hat{\beta}_2 x_{2o} + \ldots + \hat{\beta}_k x_{ko}\,.$$

Sous la supposition de normalité, nous pouvons déterminer un intervalle de confiance pour $\mu_{y.\boldsymbol{x}_o}$. Quand on détermine un intervalle de confiance pour $\mu_{y.\boldsymbol{x}_o}$, on affirme avec confiance que la *moyenne* des *Y* se situe dans les limites de l'intervalle.

La régression sert aussi à faire des *prédictions*. Une valeur future de *Y* non observée, peut être prédite à partir des valeurs des variables exogènes $\boldsymbol{x}_o = [x_{1o}, x_{2o}, \ldots, x_{ko}]$. La prédiction $\hat{y}_o$ est identique à $\hat{\mu}_{y.x_o}$:

$$\hat{y}_o = \hat{\beta}_o + \hat{\beta}_1 x_{1o} + \hat{\beta}_2 x_{2o} + \ldots + \hat{\beta}_k x_{ko}$$

Cette prédiction est bien sûr sujette à erreur. Les *limites de prédiction* sont les limites dans lesquelles devraient se trouver la prochaine *observation* sur *Y*. Ces limites de prédiction sont plus larges que celles de l'intervalle de confiance pour $\mu_{y.x_o}$.

Exemple 10.2.2 (*Suite de l'exemple* 10.2.1)
[*Nous devons utiliser ici le logiciel* **MegaStat**.]
Supposons qu'on considère une classe de maisons, celles de 2 chambres à coucher âgées de 20 ans. Voici un intervalle de confiance et des limites de prédiction à 95 % données par `MegaStats` pour les maisons de cette classe:

`Predicted values for: Prix`

			95% Confidence Interval		95% Prediction Interval	
cc	age	Predicted	lower	upper	lower	upper
2	20	48.0651	39.0761	57.0541	15.8470	80.2832

La valeur `Predicted` signifie deux choses: c'est l'estimation d'une moyenne et c'est aussi la prédiction du prix d'une maison donnée appartenant à la classe considérée.
Estimation d'une moyenne On estime que le prix moyen $\mu_{y.x}$ des maisons de la classe considérée est de 48 065 $. L'intervalle de confiance pour $\mu_{y.x}$ (`95% Confidence Interval`) permet d'affirmer avec 95 % de confiance que le prix *moyen* des maisons de 2 chambres à coucher de 20 ans d'âge se situe entre

39 076 $ et 57 054 $.

Limites de prédiction Considérons une maison de 2 chambres à coucher âgée de 20 ans à vendre. On prédit qu'elle se vendra à 48 065 $. Les limites de prédiction (`95% Prediction Interval`) permettent d'affirmer que le prix auquel cette maison se vendra se situe quelque part entre 15 847 $ et 80 283 $.

Remarque

Les limites de prédiction dans le dernier exemple sont à toutes fins pratiques inutiles, voire ridicules. Cela est dû à plusieurs facteurs:

1. Bien que la corrélation (R = 0,68) ne soit pas négligeable, elle n'est pas assez forte pour fins de prévision. Les prix varient considérablement, même lorsqu'on se limite aux maisons de même âge et ayant le même nombre de chambres à coucher. Avec cette restriction, l'écart-type des prix ($\hat{\sigma}$) est de 15 308 $—inférieur, forcément, à l'écart-type (20 295 $) des prix dans la population entière, mais assez grand quand même.
2. L'échantillon n'est pas très grand, ce qui fait que l'estimation des paramètres β_j et donc de $\mu_{y.x}$ est peu fiable, comme le montre l'intervalle 39 076 $ ≤ $\mu_{y.x}$ ≤ 57 054 $.
3. Si la *moyenne* se situe entre 39 076 $ et 57 054 $, alors le prix d'une maison donnée est d'autant plus imprévisible que les prix se dispersent par rapport à leur moyenne.
4. Finalement, il faut reconnaître que le niveau de confiance demandé (95 %) est un peu trop ambitieux. Faute de mieux, on pourrait se contenter d'un niveau de confiance inférieur. Que peut-on affirmer, par exemple, à un niveau de confiance de 50 %? Dans ce cas, les limites de prédiction sont 37 215 $ et 58 917 $.
5. Quiconque a déjà eu l'occasion de s'intéresser au prix d'une maison serait surpris du peu de succès de modèle, comme on le voit au dernier paragraphe. Il est certain qu'une agente d'immeuble expérimentée pourrait faire une prédiction plus précise avec plus de confiance. Comment ça se fait? La raison est que l'agente tiendra compte de beaucoup plus de facteurs (donc de variables exogènes) que les deux que nous avons considérés ici. Les maisons peuvent être très différentes les unes des autres, même si elles sont du même âge et ont le même nombre de chambres à coucher: elles peuvent se situer dans des rues plus ou moins prisées; elles peuvent être grandes ou petites; être bien entretenues ou pas…autant de variables exogènes qui, si elles sont introduites dans le régression, feraient réduire la variance conditionnelle (estimée à 15 308 $) et améliorer la précision des prédictions.
6. C'est un fait incontournable que la prédiction d'une *valeur donnée* (par opposition à l'estimation d'une *moyenne*) est généralement peu fiable, et que la régression ne produit pas des prédictions très précises quand on a affaire à une population naturellement très dispersée. Mais sans elle, les prédictions seraient encore moins précises.

10.3 Évaluation globale

Les tests présentés ci-dessus portent chacun sur un seul des coefficients β_j. En général, on voudra également tester la pertinence des variables exogènes prises dans leur *ensemble* ainsi que d'une mesure *globale* de la qualité de la régression (analogue au coefficient de corrélation dans le cas d'une régression simple).

L'outil principal est la décomposition présentée à la section 8.6:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

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

où SCT est la somme des carrés *totale*; SCE est la somme des carrés *expliquée*, et SCR est la somme des carrés *résiduelle*. SCT et SCR peuvent être interprétées comme des mesures d'erreur de prédiction.

- SCT mesure les erreurs commises si on prédisait *Y* sans l'apport des variables exogènes. Sans les variables exogènes, la valeur de *Y* est normalement estimée par la moyenne $\bar{y}$. L'erreur commise lorsqu'on prédit y_j par $\bar{y}$ est $(y_j - \bar{y})$ et SCT est la somme des carrés de ces erreurs.
- SCR mesure les erreurs commises lorsque y_j est prédit par $\hat{y}_i$, la prédiction basée sur le modèle de régression multiple.

On s'attend, bien sûr, à ce que SCT > SCR. La différence SCE = SCT – SCR mesure la réduction de l'erreur due à l'utilisation des variables exogènes. Si SCE est grande, c'est que la régression a permis de réaliser une bonne réduction de l'erreur.

Coefficient de corrélation multiple et coefficient de corrélation multiple ajusté

La réduction d'erreur SCE est difficile à interpréter, étant donné que sa valeur peut être grande ou petite, dépendamment de l'unité de mesure de *Y*. Cette difficulté est réglée en exprimant cette différence en termes relatifs, c'est-à-dire, comme fraction de SCT. Ce quotient est appelé *coefficient de détermination* ou *coefficient de corrélation multiple* et désigné par R^2:

$$R^2 = \textit{Coefficient de détermination} = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}$$

Il est clair que $0 \le R^2 \le 1$: $R^2 = 0$ si SCR = SCT (aucune réduction des erreurs); et $R^2 = 1$ si SCR = 0 (aucune erreur de prédiction dans the modèle de régression).

Le coefficient de corrélation multiple est simplement la racine carrée positive de R^2:

$$R = \textit{Coefficient de corrélation multiple} = \sqrt{R^2} = \sqrt{\frac{\text{SCE}}{\text{SCT}}} = \sqrt{1 - \frac{\text{SCR}}{\text{SCT}}}$$

Autre interprétation de R Une autre interprétation de *R* le réduit à la notion de coefficient de corrélation simple déjà connue. Pour mesurer la dépendance entre la variable endogène *Y* et *plusieurs* variables exogènes, $x_1, x_2, \ldots, x_k$, on remplace les *k* variables exogènes par une fonction affine de celles-ci, $\beta_o + \beta_1 x_j + \ldots + \beta_k x_k$. Comment choisir les β_j? Eh bien, on prend $\beta_j = \hat{\beta}_j$, ce qui fait que les variables exogènes sont remplacées par une seule, $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x_j + \ldots + \hat{\beta}_k x_k$, la valeur prédite de *y*. Alors

R est le coefficient de corrélation entre *y* et $\hat{y}$.

Ceci donne un deuxième sens intuitif à *R*.

Moyennes des carrés

Une somme de carrés augmente avec le nombre de termes et donc ne peut servir que si elle est ajustée pour tenir compte de ce fait, c'est-à-dire, à moins qu'on remplace la somme par une *moyenne* de carrés. Mais ce ne sera pas tout-à-fait une moyenne car on divisera non pas par le nombre de termes mais plutôt par le *nombre de degrés de liberté* (voir plus bas).

Voici le nombre de degrés de liberté des trois sommes de carrés:

Somme des carrés	Degrés de liberté	Moyennes des carrés
SCE	k	$\text{MCE} = \dfrac{\text{SCE}}{k}$
SCR	n-(k+1)	$\text{MCR} = \dfrac{\text{SCR}}{n-(k+1)}$
SCT	n-1	$\text{MCT} = \dfrac{\text{SCT}}{n-1}$

Remarque Noter que MCR est l'estimateur $\hat{\sigma}^2$ de σ^2 alors que SCT est la variance échantillonnale S_y^2 qui estime la variance des y_i sous un autre modèle, un modèle dans lequel $E(y_i) = \mu$ pour tout i.

On peut alors redéfinir R en remplaçant les sommes de carrés par les moyennes des carrés:

$$R \text{ ajusté} = \sqrt{R^2 \text{ ajusté}} = \sqrt{1 - \frac{\text{MCR}}{\text{MCT}}}$$

Remarque Il est rare que cet ajustement change la valeur de R de façon radicale. On la préfère comme mesure descriptive parce qu'elle fait payer un prix pour l'ajout de variables exogènes. L'introduction de nouvelles variables exogènes—pertinentes ou pas—ne peut que faire croître R. Le R ajusté réduit la possibilité qu'un accroissement qui ne serait dû qu'à l'ajout d'une nouvelle variable exogène non significative.

Exemple 10.3.2 (*Suite de l'exemple* 10.2.1)

MegaStat fournit R^2, R^2 ajusté ainsi que $\hat{\sigma}$ (`Std. Error`).

```
Regression Analysis
                    R²  0.458
           Adjusted R²  0.431          n  43
                    R  0.677          k  2
           Std. Error  15.308   Dep. Var.  Prix
```

Noter que $\hat{\sigma}$ n'estime pas l'écart-type de toutes les maisons; il estime l'écart-type des prix des maisons ayant toutes les mêmes valeurs des variables exogènes. L'écart-type des prix de toutes les maisons de la population peut être estimé (si les maisons ont été tirées au hasard) par S_y = 20 295 $. C'est le fait que $\hat{\sigma}$ = 15 308 $ est plus petit que S_y qui rend la régression utile.

Test global

En général, avant de s'attarder sur la significativité des coefficients β_j, il est bon d'effectuer un test global, soit un test de l'hypothèse

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0.$$

L'hypothèse affirme simplement qu'aucune des variables exogènes ne contribue à la prédiction de Y. Nous décrivons la procédure d'abord. Nous énoncerons ensuite les théorèmes qui l'appuient et nous justifierons la région critique en termes intuitifs.

Le test est basé sur la statistique $F = \dfrac{\text{MCE}}{\text{MCR}}$ et la région critique est

$$F > F_{k;n-(k+1);\,\alpha}$$

où $F_{k;n-k-1;\,\alpha}$ est le point critique correspondant à une loi de Fisher à k et n-(k+1) degrés de liberté.

Exemple 10.3.1 *Une analyse globale — décomposition d'une somme de carrés*

La commande Excel utilisée à l'exemple 10.2.1 fournit aussi le tableau suivant, une *table d'analyse de variance*:

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	5 182.7293	2	2 591.3647	8.56	.0008
Residual	12 417.9950	41	302.8779		
Total	17 600.7243	43			

Les colonnes `SS`, `df` et `MS` donnent, respectivement, les sommes de carrés SCE, SCR et SCT, le nombre de degrés de liberté et les moyennes des carrés.

La colonne *F* calcule le quotient F = MCE/MCR.

Si $\alpha = 0{,}05$, l'hypothèse H_o est rejetée si $F > \mathcal{F}_{2;41;0,05} = 3{,}22$. La valeur observée $F = 8{,}56 > 3{,}22$ entraîne donc le rejet de H_o.

La colonne *p-value* nous dispense de ces calculs. La valeur 0,0008 est la probabilité $P(F_{2;41} > 8{,}56)$, où $F_{2;41}$ est une variable de loi $\mathcal{F}$ à 2 et 42 degrés de liberté.

La conclusion est que les coefficients β_1 et β_2 ne sont pas tous nuls et donc qu'au moins l'un des deux est non nul. On peut dès lors étudier chacun des coefficients séparément.

Remarque

Il n'est pas impossible que les tests mènent au rejet de l'une des hypothèses $\beta_j = 0$ sans pour autant rejeter l'hypothèse H_o que tous les coefficients sont nuls. On ne peut pas facilement concilier ces conclusions contradictoires. En général, l'approche conservatrice est recommandée: si H_o est acceptée, on s'abstient de procéder à des tests individuels. On accepte le verdict du test global : l'utilité des variables exogènes n'est pas démontrée.

On ne peut pas, cependant, se montrer toujours aussi rigide. Il faut se rappeler lorsqu'on « accepte » H_o, on n'affirme pas qu'elle est vrai; on dit simplement qu'on ne peut pas la rejeter avec confiance. Donc si H_o est « presque » rejetée (une valeur *p* proche du seuil), il est défendable de poursuivre l'analyse avec des tests individuels sur les β_j.

Justification théorique

La région critique $F > F_{k;n-(k+1);\,\alpha}$ est basée sur les propriétés suivantes:

1. SCR ~ $\chi^2_{n-(k+1)}$
2. Sous H_o, SCE ~ χ^2_k
3. SCR et SCE sont indépendantes

Par conséquent, sous H_o, $\dfrac{\text{SCE}/k}{\text{SCR}/[n-(k+1)]} = \dfrac{\text{MCE}}{\text{MCR}} \sim \mathcal{F}_{k;n-(k+1).}$

10.4 Variables dichotomiques

En régression linéaire, la variable endogène est nécessairement quantitative. On pourrait croire que les variables exogènes sont elles aussi nécessairement quantitatives, ce qui exclurait, par exemple, qu'on tienne compte de la variable SEXE pour prédire la valeur *Y* d'une mesure d'endurance physique. Or ce n'est pas le cas, car une variable qualitative peut toujours s'exprimer par une série de variables dichotomiques, c'est-à-dire, des variables qui ne prennent que deux valeurs—comme la variable SEXE, justement. Ses valeurs (Femme/Homme) peuvent toujours être remplacées par les nombres 0 et 1.

Considérons, pour commencer, le cas d'une variable exogène dichotomique. La procédure est simple: on remplace ses deux valeurs par 0 et 1, respectivement, puis on procède comme avec une

variable quantitative ordinaire. Ce qui reste à préciser, c'est l'interprétation des résultats.

Exemple 10.4.1 *Une variable dichotomique*

Considérons les données du tableau A03, qui portent sur un échantillon de 28 personnes, dont 16 femmes et 12 hommes. On s'intéresse aux variables suivantes:

Poids	Le poids de la personne (en livres)
Taille	La taille de la personne (en pouces)
Sexe	0 = Féminin; 1 = Masculin

Nous cherchons à développer une équation permettant de prédire le poids d'une personne à partir de son sexe et sa taille. Le modèle est:

$$E(\text{Poids}) = \beta_o + \beta_1(\text{Sexe}) + \beta_2(\text{Taille})$$

Le logiciel `MegaStats` donne les estimations suivantes:

$$\hat{\beta}_o = -124; \quad \hat{\beta}_1 = 6,7356; \quad \hat{\beta}_2 = 3,9477 .$$

L'équation de prédiction serait donc

$$\textit{Poids} = -124 + 6,7357(\text{Sexe}) + 3,9477(\text{Taille}).$$

Le coefficient de Sexe est 6,7357. Que représente ce nombre? Ce nombre est la différence de poids, pour une taille donnée, entre les hommes et les femmes. Pour le voir, on énonce les résultats d'une autre façon: On présente deux équations, l'une pour les femmes et l'autre pour les hommes:

Femmes: $\textit{Poids} = -124 + 3,9477(\textit{Taille})$

Hommes: $\textit{Poids} = -124 + 6,7357 + 3,9477(\textit{Taille})$

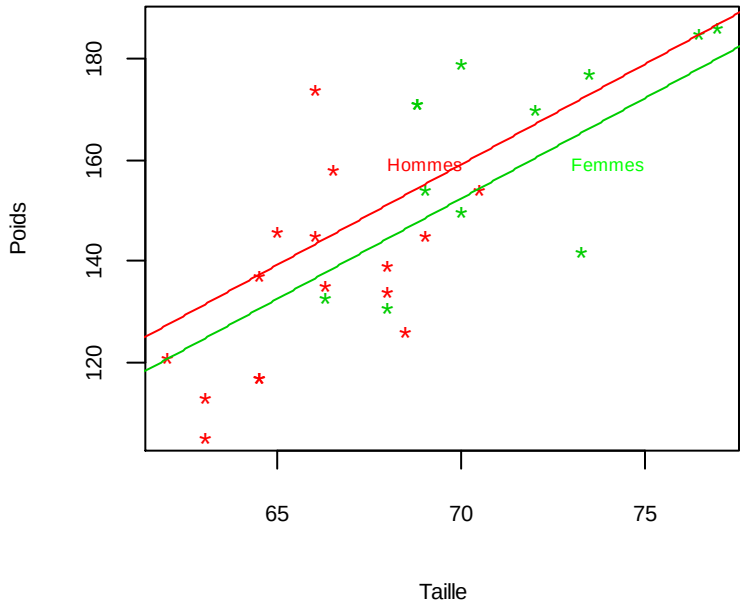
$= -117 + 3,9477(\textit{Taille})$

Graphiquement, le lien entre la taille et le poids s'exprime par deux droites, l'une pour les femmes, l'autre pour les hommes:

Figure 10.1

Relation entre la taille et le poids

Femmes et hommes



Tests d'hypothèses

Nous n'avons pas encore abordé la question de savoir si les coefficients sont significativement différents de 0. Le tableau suivant, produit par **MegaStat**, fournit les tests pertinents.

Regression output					*confidence interval*	
variables	coefficients	std. error	t (df=25)	p-value	95% lower	95% upper
Intercept	-124.0006	70.5925	-1.757	.0912	-269.3886	21.3875
Sexe	6.7357	8.0845	0.833	.4126	-9.9147	23.3860
Taille	3.9477	1.0687	3.694	.0011	1.7467	6.1487

On constate que la *valeur p* pour l'hypothèse $\beta_1 = 0$ est 0,4126. Donc l'hypothèse ne peut être rejetée à un niveau raisonnable—on ne peut pas affirmer que le sexe d'une personne, une fois sa taille connue, est utile à la prédiction du poids de la personne. Plus concrètement, le poids moyen d'une femme est le même que celui d'un homme de même taille.

On détermine alors une régression avec seule la taille comme variable exogène. Le modèle est:

$$E(Poids) = \gamma_o + \gamma_2(Taille)$$

Les sorties suivantes donnent les estimations des paramètres:

Regression output					*confidence interval*	
variables	coefficients	std. error	t (df=26)	p-value	95% lower	95% upper
Intercept	-162.3787	53.1772	-3.054	.0052	-271.6860	-53.0713
Taille	4.5531	0.7790	5.845	3.68E-06	2.9518	6.1544

Il est intéressant de noter que si seule la variable SEXE est utilisée comme variable exogène, elle se révèle hautement significative:

Regression output					*confidence interval*	
variables	coefficients	std. error	t (df=26)	p-value	95% lower	95% upper
Intercept	136.3750	4.7314	28.823	2.91E-21	126.6495	146.1005
Sexe	27.0417	7.2273	3.742	.0009	12.1856	41.8977

Question: En fin de compte, la variable SEXE est-elle, oui ou non, utile à la prédiction du poids?
Les analyses ci-dessus donnent à croire que si les femmes ont un poids inférieur à celui des hommes, c'est dû uniquement au fait qu'elles sont en moyenne plus petites de taille. (Ceci contredit ce qu'on sait des différences anatomiques entre femmes et hommes, mais rappelons que « accepter » une hypothèse n'est pas équivalent à l'affirmer.)

Dans le dernier exemple, nous avons considéré un modèle qui exprime la relation entre la taille et le poids par deux droites, l'une pour les femmes, l'autre pour les hommes. Mais la construction du modèle à contraint les deux droites à être parallèles. L'hypothèse implicite est qu'un accroissement d'une unité de taille entraîne un accroissement de poids qui est le même chez les femmes et les hommes. Il serait préférable de mettre cette hypothèse à l'épreuve car ce n'est peut-être pas le cas. La chose à faire est donc de commencer par un modèle plus général dans lequel les droites ne sont pas nécessairement parallèles et ensuite tester l'hypothèse de parallélisme. La procédure est illustrée dans le prochain exemple.

Exemple 10.4.2 [*Suite de l'exemple* 10.4.1]

Dans le modèle traité à l'exemple 10.4.1, l'équation $E(Poids) = \beta_o + \beta_1(Sexe) + \beta_2(Taille)$ implique une même pente β_2 indépendamment du sexe (alors que le terme $\beta_1(Sexe)$ qui s'ajoute seulement aux sujets masculins permet que les droites ne soient pas confondues). S'il faut distinguer hommes et femmes quant à la pente, il faudrait, selon le même principe, ajouter un terme, disons $\beta_3(Taille)$ seulement pour les sujets masculins. Pour ce faire, on utilise un artifice suivant: on crée une nouvelle

variable *SexeTaille* qui représente la taille pour les hommes mais qui prend la valeur 0 pour les femmes.

$$SexeTaille = \begin{cases} Taille & \text{si le sujet est un homme} \\ 0 & \text{sinon} \end{cases}$$

Le modèle est donc

$$E(Poids) = \beta_o + \beta_1(Sexe) + \beta_2(Taille) + \beta_3(TailleSexe)$$

Regression output

variables	coefficients	std. error	t (df=24)	p-value
Intercept	-143.2724	112.9266	-1.269	.2167
Sexe	40.1774	151.2235	0.266	.7928
Taille	4.2399	1.7111	2.478	.0206
SexeTaille	-0.4915	2.2192	-0.221	.8266

On résume

Femmes: $Poids = -143{,}2724 + 4{,}2399(Taille)$

Hommes: $Poids = (-143{,}2724+40{,}1774) + (4{,}2399-0{,}4915)(Taille)$

$= -103{,}0950 + 3{,}7484(Taille)$

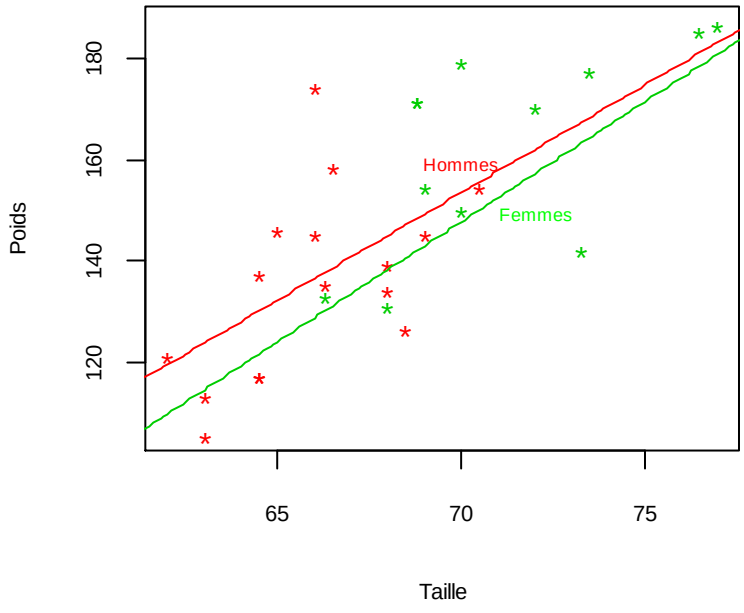
La première hypothèse à tester est $\beta_3 = 0$: c'est l'hypothèse que les deux droites sont parallèles. On est loin de pouvoir la rejeter, et on l'exclura du modèle pour revenir au modèle considéré dans le dernier exemple.

La figure 10.2 illustre la relation.

Figure 10.2

Relation entre la taille et le poids

Femmes et hommes



Remarque

Nous n'avons pas rejeté l'hypothèse $\beta_3 = 0$ et nous l'avons donc exclu du modèle, c'est-à-dire que le terme en β_3 n'entrera pas dans notre description de la relation entre la taille et le poids, et ne servira pas à la prédiction poids. Mais cette décision n'est pas automatique. Si on exclut le terme ici, c'est d'abord parce que la *valeur p* est très élevée (0,8266), mais aussi parce le signe (négatif) de β_3 est a priori peu crédible.

Dans d'autres circonstances, on pourrait très bien conserver un terme qui n'est pas significativement différent de 0—c'est ce qu'on aurait fait ici, par exemple, si $\hat{\beta}_3$ avait été positif et si la *valeur p* avait été proche du seuil.

10.5 Régression polynomiale

Évidemment, la relation entre deux variables n'est souvent pas linéaire. La figure 10.3, qui illustre la relation entre la grosseur d'un diamant et son prix en est un exemple. le nuage de points présente une courbure qui montre que le modèle $E(Y) = \beta_o + \beta_1 x$ qui impose une structure linéaire s'ajuste mal aux données. Il faut alors songer à une fonction non linéaire. Parmi celles-là se trouvent les fonctions polynomiales. Par exemple, $E(Y) = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. Ce modèle est un modèle de régression multiple, avec x, x^2 et x^3 comme variables exogènes.

Exemple 10.5.1 *Régression polynomiale*

Le tableau 10.3 présente des données sur le prix et la grosseur d'un échantillon de diamants. La figure 10.2 montre bien que la relation n'est pas linéaire. Considérons donc le modèle $E(Prix) = \beta_o + \beta_1(Carat) + \beta_2(Carat2)$, où $Carat2 = Carat^2$.

Voici les résultats fournis par `MegaStats`:

`Regression output`

					confidence interval	
variables	coeffi-cients	std. error	t (df=97)	p-value	95% lower	95% upper
Intercept	259.2403	466.9891	0.555	.5801	-667.6038	1 186.0843
Carats	1 676.8083	1 694.8114	0.989	.3249	-1 686.9232	5 040.5399
Carats2	7 647.7129	1 346.6735	5.679	1.41E-07	4 974.9388	10 320.4870

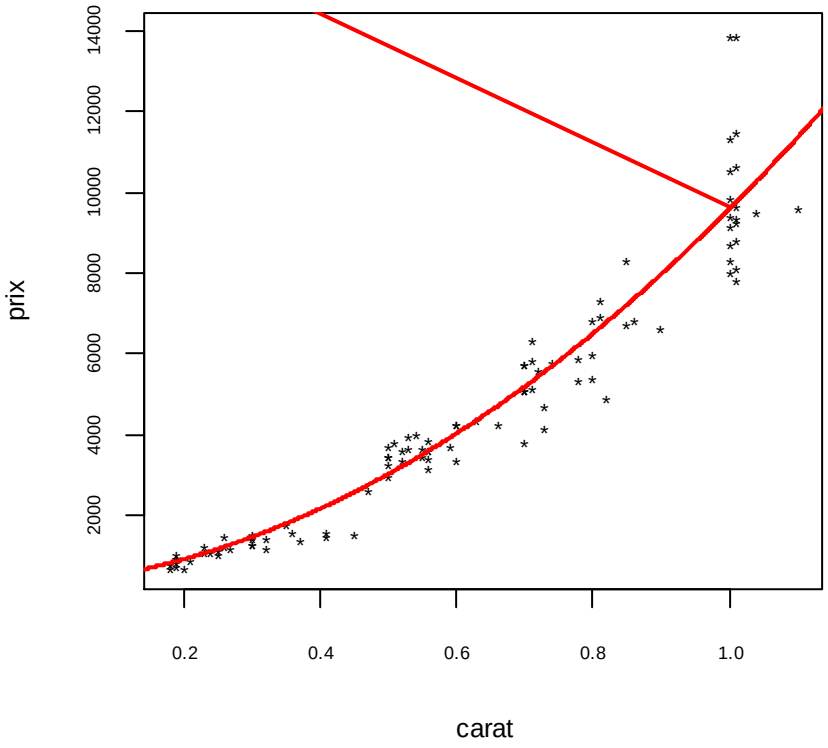
On estime donc que $E(Prix) = 259 + 1\,677(Carats) + 7\,648(Carats^2)$. On constate que le terme Carats2 est fortement significatif, ce qui confirme qu'une droite n'est pas suffisante pour exprimer la relation entre le prix et la grosseur (un fait bien connu dans le domaine, d'ailleurs).

Le terme linéaire $\hat{\beta}_1$ (le coefficient de *Carats*) est tout à fait non significatif. Peut-on l'éliminer du modèle? Normalement, le terme linéaire est conservé, car même si $\hat{\beta}_1$ est non significatif, il n'est pas prouvé que $\beta_1 = 0$. Dans le cas présent on pourrait l'omettre il ne contribue presque rien à R^2: il le fait passer de 0,921 à 0,922. On peut alors proposer comme modèle définitif un modèle avec *Carats*2 comme seul variable exogène.

Regression output					confidence interval	
variables	coefficients	std. error	t (df=98)	p-value	95% lower	95% upper
Intercept	696.4480	150.9855	4.613	1.20E-05	396.8222	996.0738
Carats2	8 954.0316	264.9640	33.793	8.23E-56	8 428.2192	9 479.8441

L'équation définitive serait alors $E(Prix) = 696 + 8954(Carats2)$

Figure 10.3
Prix de 100 diamants
en fonctioin de leur grosseur



10.6 Analyse de variance par la régression

L'analyse de variance à un facteur, traitée au chapitre 9 peut également être considéré comme un cas particulier d'une régression multiple, car il est possible d'exprimer les valeurs d'une variable qualitative à l'aide d'un certain nombre de variables dichotomiques. L'exemple suivant illustre la façon de faire

Exemple 10.6.1 *Une analyse de variance effectuée via une régression multiple*

Les données suivantes sont tirées de l'exercice 9.10. Ce sont les prix (en milliers) ses maisons d'une ou deux chambres à coucher vendues dans trois secteurs d'une ville:

Secteur Sud:	249	184	239	314	69	69	80	85	85	269	60	89	89	89	89	142	142
Secteur Nord:	97	157	170	184	385	145											
Secteur Centre:	98	249	289	299													

L'objectif de l'analyse est de savoir s'il y a une différence entre les trois secteurs quant au prix moyen des maisons. Pour effectuer une régression avec *Prix* pour variable endogène et *Secteur* pour variable exogène, nous devons exprimer cette dernière par des variables quantitatives. On commence par prendre l'un des secteurs comme base de comparaison. Ce pourrait être n'importe lequel. Choisissons le secteur sud. Nous définirons alors deux variables exogènes:

$$Nord = \begin{cases} 1 & \text{pour une maison dans le secteur nord} \\ 0 & \text{pour une maison ailleurs} \end{cases}; \qquad Centre = \begin{cases} 1 & \text{pour une maison dans le secteur centre} \\ 0 & \text{pour une maison ailleurs} \end{cases}$$

Les valeurs des variables *Prix*, *Nord* et *Centre* sont présentées au tableau 10.5.1.

Le modèle est

$$E(Prix) = \beta_o + \beta_1(Nord) + \beta_2(Centre)$$

Le secteur étant le secteur sud, les paramètres β_1 et β_2 représentent l'écart moyen des prix des maisons des secteurs nord et centre, respectivement, par rapport à ceux du secteur sud.

Le logiciel MegsStats produit les estimations des paramètres:

Regression output					confidence interval	
variables	*coefficients*	*std.error*	*t (df=24)*	*p-value*	*95% lower*	*95% upper*
Intercept	137.8235	21.4144	6.436	1.18E-06	93.6263	182.0207
Nord	60.7765	44.9193	1.353	.1887	-31.9324	153.4853
Centre	78.1765	44.9193	1.740	.0946	-14.5324	170.8853

On a $\hat{\beta}_1 = 60{,}7765$ et $\hat{\beta}_2 = 78{,}1765$. On estime donc que les maisons des secteurs nord et centre valent en moyenne 60 777 $ et 78 176 $ de plus que celles du secteur sud.

Les *valeurs p* = 0,1887 et *p* = 0,0946 testent les hypothèses $\beta_1 = 0$ et $\beta_2 = 0$, respectivement. On ne peut pas raisonnablement rejeter la première hypothèse; la deuxième pourrait l'être, si on adopte un niveau $\alpha = 0{,}1$.

On peut réunir les deux hypothèses en une seule hypothèse H_o qui combine les deux, soit:

$$H_o : \beta_1 = 0 \text{ et } \beta_2 = 0,$$

ce qui revient à dire que le prix d'une maison ne dépend pas du secteur.

C'est précisément ce que teste l'analyse de variance (ANOVA) fournie par MegaStats:

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	31 148.9961	2	15 574.4980	2.00	.1576
Residual	187 099.6706	24	7 795.8196		
Total	218 248.6667	26			

Puisque *p* = 0,1576, on ne rejette pas H_o.

Cette analyse est tout à fait équivalente à celle qu'on aurait faite au chapitre 9.

10.7 Exemples artificiels

Cette section a pour but d'illustrer comment le lien entre deux variables peut être influencé par une troisième. Dans le premier exemple, une dépendance entre deux variables est estompée sous l'effet d'une troisième. Dans le deuxième exemple, au contraire, une dépendance est créée artificiellement par une troisième variable. Les deux exemples qui suivent—fictifs et caricaturaux—sont conçus pour mettre le phénomène en évidence.

Exemple 10.7.1 *Une corrélation éliminée sous l'effet d'une troisième variable*

Dans la figure 10.4 présente les données du tableau 10.5, l'axe vertical est le score dans un test de vocabulaire et l'axe horizontal est l'âge. On ne décèle aucune dépendance entre les deux.

Un modèle qui ne tient compte que de l'âge, soit $E(voc) = \beta_o + \beta_1(age)$, on trouve $\hat{\beta}_o = 49{,}65$ et $\hat{\beta}_1 = 0{,}0037$, ce qui donne l'équation suivante:

E(voc) = 49,65 + 0,0037(age).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.648938   0.549067  90.424   <2e-16 ***
age          0.003732   0.011286   0.331    0.742    Coefficients:
Multiple R-squared:  0.01196,   Adjusted R-squared:  -0.008625
F-statistic: 0.581 on 1 and 48 DF,  p-value: 0.4496
```

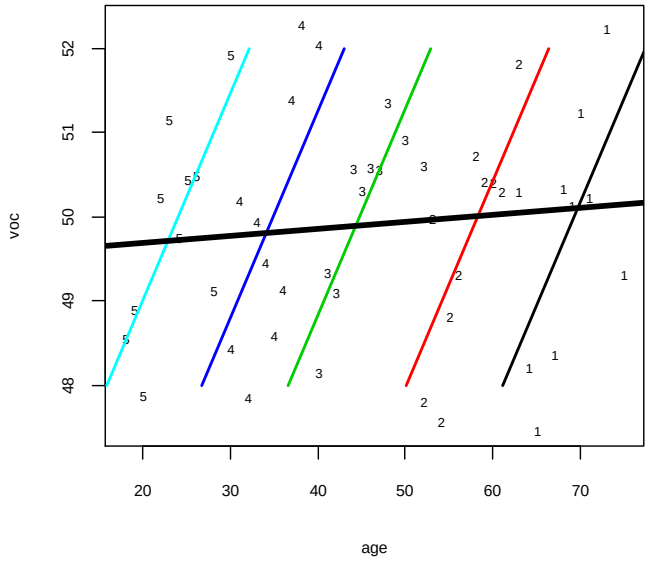
Mais un r^2 de 0,012 et une *valeur p* de 0,742 ne permet pas de conclure que $\beta_1 \neq 0$. Conclusion provisoire: pas de relation entre l'âge et le vocabulaire.

Pourtant, le vocabulaire normalement augmente avec l'âge. Ce qui détruit en apparence cette dépendance, c'est une autre relation, celle entre l'âge et la scolarité: il se trouve que, dans la population, les personnes âgées sont moins scolarisées que les plus jeunes, et de ce fait ont un vocabulaire plus faible—plus faible que ne laisserait prévoir leur âge. L'effet positif de l'âge est annulé par l'effet négatif d'une faible scolarisation.

Dans le graphique, les points du nuage sont remplacés par des chiffres de 1 à 5 représentent des niveaux de scolarité: 1 pour le plus bas niveau, 5 plus le plus haut. On constate que le plus haut niveau de scolarité (5) se trouve parmi les plus jeunes.
On remarque que si on se limite à une scolarité donnée (par exemple, aux seuls points identifiés par «5», on décèle une relation très nette entre le vocabulaire et la scolarité—et ce pour tout niveau de scolarité.
On définit alors la variable scol, le niveau de scolarité, et on l'introduit dans le modèle

$$E(\text{voc}) = \gamma_o + \gamma_{1(}\text{age}) + \gamma_2(\text{scol}).$$

Figure 10.4
Relation entre le vocabulaire et l'âge
selon le niveau de scolarité




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.32895    3.10354   9.772 6.70e-13 ***
scol         2.80688    0.44696   6.280 1.01e-07 ***
age          0.24150    0.03878   6.227 1.22e-07 ***
Residual standard error: 0.9692 on 47 degrees of freedom
Multiple R-squared:  0.4575,    Adjusted R-squared:  0.4344
F-statistic: 19.82 on 2 and 47 DF,  p-value: 5.735e-07
```

On obtient les estimations $\hat{\gamma}_o = 30{,}32$; $\hat{\gamma}_1 = 0{,}2415$; et $\hat{\gamma}_2 = 2{,}807$.

L'équation est donc la suivante:

$$E(\textit{voc}) = 30{,}33 + 0{,}2415(\textit{age}) + 2{,}807(\textit{scol}).$$

Les tests sur γ_1 et γ_2 permettent de conclure que $\gamma_1 \neq 0$ ($p = 1{,}22\times10^{-7}$) et que $\gamma_2 \neq 0$ ($p = 1{,}01 \times 10^{-7}$). Le coefficient $\hat{\gamma}_1 = 0{,}2415$ représente l'accroissement du score de vocabulaire par année d'âge pour *des personnes de même scolarité.*

Remarque

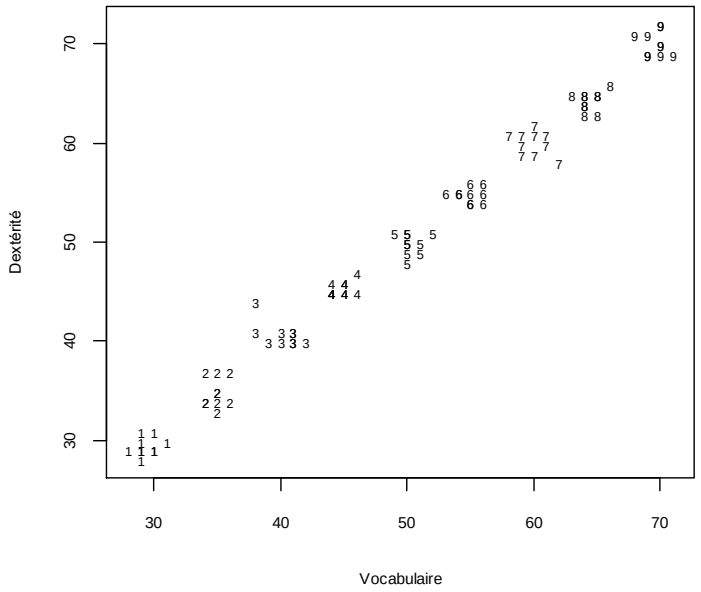
Dans le dernier exemple, une dépendance réelle et positive (entre le vocabulaire et l'âge) est effacée sous l'effet d'une troisième variable, la scolarité. L'exemple 10.7.1 illustre la différence entre une corrélation ordinaire et ce qu'on appelle une *corrélation partielle*. La corrélation entre le vocabulaire et l'âge est sujette à l'action d'autres variables, dont la scolarité. Dans le cas de l'exemple, cette action a pour effet de réduire la corrélation à presque rien. La corrélation partielle *étant donné la scolarité* tente d'éliminer l'effet de cette troisième variable. C'est la corrélation qu'on aurait calculée si on pouvait se limiter à des individus ayant tous le même niveau de scolarité. On ne peut pas entrer ici dans le détail de ce calcul, mais le concept de corrélation partielle est révélé dans l'exemple: on voit nettement que si on se limite à un niveau de scolarité donné, la corrélation entre le vocabulaire et l'âge est forte. C'est le coefficient de corrélation partielle.

L'inverse de ce qu'illustre l'exemple 10.7.1 est possible aussi: une corrélation entre deux variables qui n'est due qu'à l'action d'une troisième variable. Ce phénomène est illustré dans l'exemple suivant.

Exemple 10.7.2 *Une corrélation créée sous l'effet d'une troisième variable*

La figure 10.5 présente la relation entre la dextérité manuelle (en ordonnée) et le score en un test de vocabulaire (en abscisse) (Tableau 10.6). On ne décèle aucune dépendance entre les deux.

Figure 10.5
Relation entre la déxtérité manuelle et le vocabulaire de 90 enfants compte tenu du groupe d'âge



Les chiffres dans le graphique représentent les groupes d'âge, « 1 » étant les plus jeunes, « 9 » les plus âgés. Globalement, la relation entre la dextérité et le vocabulaire est positive et prononcée, alors que, dans un groupe donné elle, de toute évidence, inexistante.

Considérons un modèle qui ne tient pas compte de l'âge, soit le modèle E(*Dext*) = $\beta_o + \beta_1$(*voc*), où la dextérité et le vocabulaire sont désignés par *Dext* et *voc*, *respectivement.* On obtient les estimations $\hat{\beta}_o$ = 0,4506 et $\hat{\beta}_1$ = 0,9961. La relation est donc E(*Dext*) = 0,4506 + 0,9961(*voc*).

Pour tenir compte de l'âge, on l'introduit la variable *groupe* dans la régression, le modèle étant E(*Dext*) = $\gamma_o + \gamma_1$(*voc*) + γ_2(*groupe*). On trouve $\hat{\gamma}_o$ = 25,70308, $\hat{\gamma}_1$ = -0,02389; $\hat{\gamma}_2$ = 5,11072. La relation est donc E(*Dext*) = 25,70308 - 0,02389(*voc*) + 5,11072(*groupe*). On trouve que γ_1 est possiblement nul (p = 0,678) alors que γ_2 est certainement positif ($p = 2{,}2\times10^{-16}$).