CHAPITRE 12

Les régressions et corrélations multiples

Au chapitre 10, nous avons examiné les techniques de régression et de corrélation décrivant la relation entre deux variables mesurées par des échelles d'intervalles ou de proportion. Ce chapitre-ci présente un prolongement de ces techniques bivariées en introduisant des variables additionnelles dans l'analyse. Une fois que nous aurons étendu ainsi le modèle de régression bivariée à l'analyse multivariée, nous nous attarderons aux coefficients de corrélation partielle qui mesurent l'intensité et la direction d'une relation tout en contrôlant l'effet d'une ou de plusieurs variables supplémentaires. Ensuite nous examinerons le coefficient de corrélation multiple, une statistique qui mesure l'effet simultané de deux variables indépendantes ou plus sur une variable dépendante. En cours de route, nous étudierons le test de signification des coefficients de corrélation partielle et de corrélation multiple et nous apprendrons à créer une variable factice de manière à pouvoir appliquer l'analyse de corrélation à des variables nominales.

À la fin de ce chapitre, vous pourrez :

1. Comprendre l'analyse de régression à deux variables indépendantes ou plus.
2. Reconnaître les conditions d'application des méthodes de corrélation partielle et de régression multiple.
3. Calculer et interpréter des coefficients de corrélation multiple.
4. Calculer et interpréter des coefficients bêta.
5. Effectuer et interpréter des tests de signification pour des corrélations partielles et multiples.
6. Expliquer et créer des variables factices.

12.1 L'extension du modèle de régression

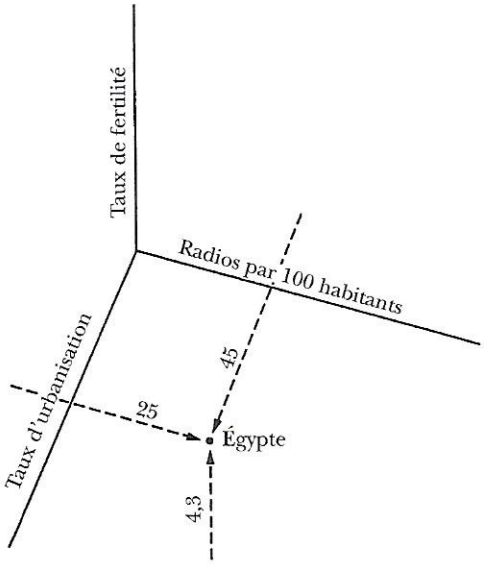
Nous avons vu dans le chapitre 10 (section 10.12) que, pour les 50 pays les plus peuplés du monde, il existe une forte corrélation entre le taux de fertilité et à la fois le taux d'urbanisation et le nombre de radios par 100 habitants. Le pourcentage de la population habitant en ville explique 31 % de la variation du taux de fertilité ($r^2 = 0,31$) et le nombre de radios par 100 habitants explique 22 % de la variation du taux de fertilité. Mais quel est le pourcentage du taux de fertilité qui est expliqué par la conjugaison du taux d'urbanisation et du nombre de radios ? On ne peut pas tout simplement additionner les pourcentages individuels de la variation expliquée par l'urbanisation et le nombre de radios car ces deux variables indépendantes sont elles-mêmes corrélées ($r = 0,62$). Leurs effets sur le taux de fertilité se chevauchent donc, et il nous faut trouver un moyen pour tenir compte de ce chevauchement lorsque nous mesurons l'effet combiné de l'urbanisation et du nombre de radios. Il nous faut également trouver comment mesurer l'effet de chaque variable indépendante sur la variable dépendante tout en contrôlant l'effet de l'autre variable indépendante.

Nous pouvons mesurer les effets de deux variables indépendantes ou plus sur une variable dépendante en étendant le modèle de régression linéaire présenté au chapitre 10. Situons les points de données de nos trois variables dans un diagramme à trois dimensions, avec les variables indépendantes (taux d'urbanisation et nombre de radios par 100 habitants) représentées par deux axes et la variable dépendante (taux de fertilité) représentée par un troisième axe. Comme on le fait pour les diagrammes de dispersion bivariés, on place la variable dépendante sur l'axe des Y. Ce diagramme de dispersion est assez facile à visualiser. Pensez à un coin dans une pièce. Le plancher est le diagramme de dispersion pour la relation entre le taux d'urbanisation et le nombre de radios. Un des murs représente la relation entre le nombre de radios et le taux de fertilité. Le dernier mur est consacré à la relation entre le taux d'urbanisation et le nombre de radios. On situe chaque cas dans l'espace à trois dimensions à l'intersection des scores du cas pour les trois variables.

Prenons l'exemple de l'Égypte. Les données pour les nations les plus peuplées indiquent que 45 % des Égyptiens vivent en ville, qu'il y a 25 radios par 100 Égyptiens et que les femmes égyptiennes ont en moyenne 4,35 enfants. Nous situons donc le point pour l'Égypte à 45 unités sur l'axe de l'urbanisation, 25 sur l'axe des radios et 4,35 sur l'axe de la fertilité. La figure 12.1 montre la position de l'Égypte à l'intérieur d'un diagramme à trois dimensions que l'on pourrait

imaginer comme le coin d'une chambre. Imaginez que nous situions de cette façon les 50 cas dans cet espace tridimensionnel. Cela rendrait le diagramme confus, je ne le tracerai donc pas ici. Notre esprit travaille mieux qu'une feuille à deux dimensions lorsqu'il s'agit de représenter des diagrammes en trois dimensions. Je pense donc que vous pourrez aisément imaginer un tel diagramme de dispersion pour les 49 pays les plus peuplés. Quarante-neuf plutôt que 50 parce que nous avons dû exclure un pays – l'Ouzbékistan – pour lequel nous n'avons pas de données quant au nombre de radios. L'exclusion de l'Ouzbékistan se fera « en liste » dans l'ensemble de l'analyse.

Figure 12.1. Diagramme de dispersion à trois dimensions (Égypte seulement)



Il nous faut maintenant une méthode pratique qui résumerait les relations illustrées par le diagramme de dispersion. Repensez au modèle de régression bivariée des sections 10.1 et 10.2. Nous avons alors résumé la relation bivariée à l'aide d'une droite de régression représentée par cette équation :

$$Y = a + bX$$

Lorsque Y = le score de la variable dépendante

a = l'intersection, ou le point où la droite de régression coupe l'axe des Y

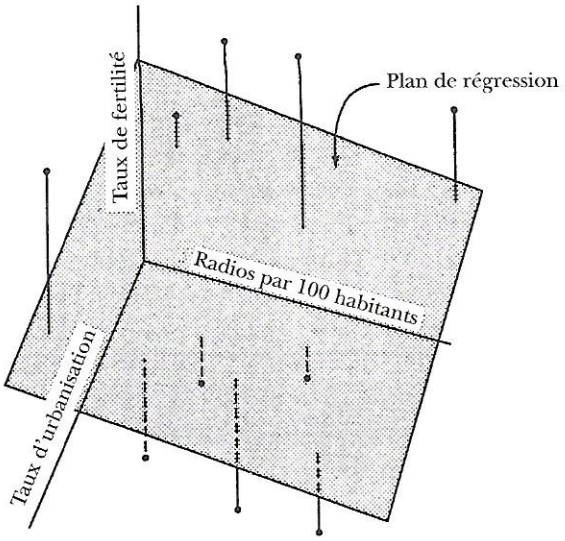
b = la pente, ou le changement en Y correspondant à un changement d'une unité en X

X = le score de la variable indépendante

Cette droite de régression constituerait le meilleur résumé des scores d'un diagramme de dispersion dans la mesure où elle minimiserait la somme des carrés des écarts entre les scores réels de la variable dépendante et les scores prédits par cette droite de régression.

Étendons maintenant la notion de droite de régression à la régression d'une relation entre une variable dépendante et *deux* variables indépendantes. De la même façon que nous avons tracé dans le diagramme une droite qui représentait le meilleur résumé de la relation, nous pouvons tracer un plan à deux dimensions dans cet espace tridimensionnel. Le plan traverse le nuage de points de telle façon qu'il minimise la somme des carrés des distances (dans la direction verticale, celle de la variable dépendante) entre chaque point et le plan. Schématiquement, ce plan des moindres carrés ressemblerait au plan traversant le diagramme tridimensionnel de la figure 12.2.

Figure 12.2. Plan de régression dans un diagramme à trois dimensions



J'ai situé 10 points dans la figure 12.2, cinq au-dessus du plan de régression et cinq au-dessous. J'ai aussi tracé des lignes verticales qui indiquent la distance entre chaque point et le plan de régression. Ces distances sont les résidus, ou les différences entre le score réel du taux de fertilité et le score de cette variable qui est prédit par le plan

de régression. Nous avons déjà parlé des résidus dans le cadre d'un diagramme en deux dimensions pour une relation bivariée. Maintenant nous les retrouvons dans un diagramme à trois dimensions pour une relation multivariée. Comme dans les relations bivariées, ces résidus sont des erreurs – la différence entre le score réel et le score prédit par le plan de régression.

De la même façon que nous représentons la droite de régression par l'équation $Y = a + bX$, nous pouvons représenter ce plan multivarié à l'aide de cette équation de régression multiple :

$$Y = a + b_1X_1 + b_2X_2$$

lorsque Y = score prédit de la variable Y

a = l'intersection, c'est-à-dire le point où le plan coupe l'axe de la variable dépendante

b_1 = la pente du plan par rapport à la variable indépendante X_1

X_1 = le score de la variable indépendante X_1

b_2 = la pente du plan par rapport à la variable indépendante X_2

X_2 = le score de la variable indépendante X_2

Dans notre exemple, Y représente le taux de fertilité (la variable dépendante) et X_1 et X_2 représentent respectivement le taux d'urbanisation et le nombre de radios pour 100 habitants. Cette équation peut être employée pour prédire le taux de fertilité à partir de ce que nous savons à propos du taux d'urbanisation dans chacun des 50 pays les plus peuplés. Les coefficients b_1 et b_2 sont des ***coefficients de régression non standardisés.*** L'adjectif non standardisé distingue ces coefficients des coefficients standardisés que nous verrons plus tard dans ce chapitre, bien que les chercheurs les appellent plus simplement des coefficients de régression si le contexte permet de voir clairement qu'on fait référence à ce type de régression. Les coefficients de régression non standardisés sont des pentes partielles qui décrivent le changement dans la variable dépendante Y associé à une augmentation d'une unité dans la variable indépendante X tout en contrôlant *l'effet de l'autre variable indépendante.*

Nous calculons l'intersection « a » (qu'on appelle parfois la constante) et le coefficient de régression non standardisé b_1 à partir des valeurs des coefficients de régression d'ordre zéro. Les formules sont cependant un peu compliquées et nous les laisserons de côté. Nous nous servirons de l'ordinateur pour les trouver. En fait, c'est ce que j'ai fait et j'ai trouvé cette équation qui décrit la régression de la

relation entre le taux de fertilité, le taux d'urbanisation et le nombre de radio pour 100 habitants.

$$Y' = 5{,}59 - 0{,}032X_1 - 0{,}010X_2$$

lorsque Y' = taux de fertilité prédit

X_1 = taux d'urbanisation

X_2 = nombre de radios pour 100 habitants

Le signe négatif du coefficient de régression indique que l'urbanisation et la possession de radios réduisent le taux de fertilité. Plus il y a d'urbanisation, plus il y a de radios, plus le taux de fertilité est bas. La grandeur des coefficients de régression non standardisés indique l'importance de l'effet de chaque variable sur le taux de fertilité tout en contrôlant l'effet de l'autre variable indépendante. Ainsi, si on contrôle l'effet du nombre de radios, une augmentation d'un point de pourcentage du taux d'urbanisation entraîne une diminution du taux de fertilité de 0,032 enfant par femmes. De la même façon, en contrôlant l'effet de l'urbanisation, une augmentation de 1 radio par 100 habitants entraîne une baisse du taux de fertilité de 0,010 enfant par femme.

L'équation de régression peut être utilisée pour prédire les scores de la variable dépendante – le taux de fertilité dans notre exemple. Nous pouvons par exemple insérer les scores d'urbanisation et de nombre de radios pour l'Égypte dans l'équation de régression multiple :

$$\begin{aligned} Y' &= 5{,}59 - 0{,}032X_1 - 0{,}010X_2 \\ &= 5{,}59 - 0{,}032(45) - 0{,}010(25) \\ &= 5{,}59 - 1{,}44 - 0{,}25 \\ &= 3{,}90 \end{aligned}$$

Comme pour les régressions bivariées, l'apostrophe du Y' nous rappelle que nous prédisons le taux de fertilité. Nos informations sur le taux d'urbanisation et le nombre de radios pour 100 habitants de l'Égypte permettent de prédire un taux de fertilité de 3,90 pour ce pays. En fait, l'Égypte a un taux de fertilité de 4,35. Le taux réel de fertilité de l'Égypte est donc supérieur de 0,45 au taux prédit à partir du taux d'urbanisation et du nombre de radios (4,35 – 3,90 = 0,45). Notre équation de régression nous mène donc à commettre cette petite erreur dans la prédiction du taux de fertilité de l'Égypte. Nous avons néanmoins amélioré nos prédictions en utilisant cette information à propos des variables indépendantes. Sans information sur le taux d'urbanisation et le nombre de radios par 100 habitants de

l'Égypte, notre meilleure prédiction du taux de fertilité serait de 3,56, la moyenne pour l'ensemble des 49 pays. Le taux réel de fertilité est supérieur de 0,79 à la prédiction basée sur la moyenne (4,35 – 3,56 = 0,79).

Il n'y a aucun nouveau concept statistique dans tout cela, simplement le prolongement de « vieilles » idées empruntées à la régression bivariée. Cependant, en plus de la condition exigeant un niveau de mesure d'intervalles ou de proportion – une condition qui vaut également pour tous les modèles de régression examinés dans ce manuel –, il y a quelques autres conditions qui doivent être remplies si l'on veut procéder à une régression multiple. Comme dans le cas de la régression bivariée, nous postulons que les variables indépendantes sont liées de façon linéaire à la variable dépendante. Après tout, un plan plat est l'équivalent bidimensionnel d'une droite unidimensionnelle. Si les relations entre les variables indépendantes et la variable dépendante ne sont pas linéaires, ce modèle décrira bien mal la relation réelle. Un plan de régression, par exemple, ne peut pas décrire une relation curvilinéaire en forme de selle, pas plus qu'une droite ne peut décrire une relation bivariée en forme de U. Les mesures-sommaires basées sur un modèle linéaire représenteront mal une telle relation et sous-estimeront son intensité. Comme dans le cas des régressions bivariées, les relations curvilinéaires peuvent être « redressées » en transformant les variables. Mais les techniques qui permettent cela dépassent l'objet de ce manuel.

Le second postulat sur lequel repose le modèle de régression multiple est moins évident : les effets des variables indépendantes sur la variable dépendante sont additifs, sans ***interaction statistique*** entre eux. Nous postulons, par exemple, que, bien que le taux d'urbanisation et le nombre de radios pour 100 habitants puissent affecter le taux de fertilité, il n'y a pas d'effet combiné du taux d'urbanisation et du nombre de radios. Les pays qui ont un haut taux d'urbanisation et un grand nombre de radios peuvent avoir un taux de fertilité plus bas, mais nous postulons que la combinaison particulière d'un haut taux d'urbanisation et d'un grand nombre de radios n'a aucun effet supplémentaire sur le taux de fertilité.

Le troisième postulat est, lui aussi, loin d'être évident : nous postulons que les variables indépendantes dans le modèle ne sont pas corrélées entre elles. Il est rare que cette condition soit entièrement respectée (et elle ne l'est pas dans notre exemple). Heureusement, le modèle de régression multiple est assez robuste pour permettre une certaine corrélation entre les variables indépendantes sans affecter sérieusement les conclusions qu'on peut tirer de l'analyse. Mais il faut être prudent lorsque les corrélations entre les

variables indépendantes sont très fortes – les statisticiens appliquent à cette situation le terme ***multicolinéarité***. Il est très difficile, voire impossible, d'isoler les effets de variables indépendantes qui sont fortement corrélées. Bien sûr, il est possible de vérifier, à l'aide de techniques de corrélation bivariée, si cette condition est respectée.

En résumé, nous supposons des variables d'intervalles/ratio, des relations linéaires, les effets additifs des variables indépendantes (sans interaction) et des variables indépendantes qui ne sont pas fortement corrélées. Voilà qui n'est pas une mince affaire ! Et nous ajouterons d'autres conditions lorsque nous nous pencherons sur les tests de signification pour les coefficients de corrélation partielle et les coefficients de corrélation multiple.

12.2 Le coefficient de corrélation multiple

Mais nous ne verrons les tests de signification que plus loin dans ce chapitre. Nous devons d'abord voir ce qu'est ***le coefficient de corrélation multiple***, une mesure de l'effet combiné d'un ensemble de variables sur une variable dépendante. On peut par exemple utiliser un coefficient de corrélation multiple pour évaluer l'effet combiné de l'urbanisation et du nombre de radios sur le taux de fertilité dans les pays les plus peuplés. En d'autres mots, le coefficient de corrélation multiple, similaire au r des relations bivariées, indique la distance des points de données par rapport au plan de régression. Le coefficient de corrélation multiple est représenté par le symbole $R_{Y \bullet 12...}$, lorsque Y représente la variable dépendante, et les indices numériques qui suivent le point (•) représentent les variables indépendantes X_1, X_2... Les indices numériques sont généralement omis lorsque l'identité des variables indépendantes est claire dans le contexte.

Comme toutes les mesures d'association « qui se respectent », R est égal à 1 (et jamais supérieur à 1) lorsque l'association est parfaite. Contrairement à un coefficient de corrélation bivariée ou partielle, le coefficient de corrélation multiple est toujours positif. Un coefficient de corrélation multiple négatif n'aurait pas de sens car nous ne pouvons pas parler de la direction d'une relation impliquant plus d'une variable indépendante. Certaines variables indépendantes peuvent avoir une relation positive avec la variable dépendante, d'autres peuvent avoir une relation négative avec la même variable dépendante. Par la corrélation multiple, nous cherchons à mesurer l'intensité de leurs effets combinés. Donc, R varie de 0 à 1,00.

Comme pour les autres types de coefficients de corrélation, le carré du coefficient de corrélation multiple ($R^2_{Y \bullet 12...}$) exprime la proportion de la variation dans la variable dépendante expliquée par

l'ensemble des variables indépendantes $X_1, X_2, \ldots$ $R^2_{Y \cdot 12 \ldots}$ s'appelle le ***coefficient de détermination multiple***, mais on l'appelle généralement ***le R-carré***. La proportion de la variation de la variable dépendante qui n'est pas expliquée par les variables indépendantes est égale à $1 - R^2_{Y \cdot 12 \ldots}$.

Au chapitre 4, nous avons appris à exprimer la variation par la somme des carrés. La variation totale d'une variable dépendante Y est mesurée par la variation par rapport à la moyenne, et peut donc être exprimée par $\Sigma(Y_i - \overline{Y})^2$. La somme des carrés expliquée par la régression de la relation entre une variable dépendante et un groupe de variables indépendantes mesure la variation des scores prédits par rapport à la moyenne – qu'on exprime aisément par $\Sigma(Y' - \overline{Y})^2$ lorsque Y' représente les scores prédits par l'équation de régression multiple. Nous avons trouvé le score prédit pour l'Égypte – 3,90. Maintenant, imaginez que l'on prédise les scores des 49 pays, qu'on en soustraie la moyenne générale, que l'on mette au carré les différences ainsi obtenues et qu'on les additionne. Cette somme de carrés mesurerait la variation de la variable dépendante expliquée par les variables indépendantes. La variation qui ne serait pas expliquée – les résidus – est basée sur les différences entre les scores réels et les scores prédits et est exprimé par $\Sigma(Y_i - Y')^2$. Oui, c'est de nouveau une somme de carrés.

Cela devrait vous sembler familier. Notez l'analogie directe avec la somme des carrés utilisée dans l'analyse de variance. En fait, on peut résumer ces sommes de carrés de régression multiple dans un tableau similaire à un tableau d'ANOVA. Trouver des sommes de carrés est assez ennuyant et nous ne nous préoccuperons pas de les calculer ici. Par chance, les ordinateurs calculent aisément ces sommes de carrés pour nous. Voici donc les sommes de carrés de notre exemple :

Source	Somme des carrés	dl	Somme des carrés moyenne	*F*	p
Régression	56,527	2	28,263	11,591	0,001
Résidus	112,167	46	2,438		
Total	168,694	48			

La somme des carrés de la régression est la somme des carrés expliquée par le groupe des variables indépendantes. Puisque R^2 est la proportion de la variance de la variable dépendante expliquée par les variables indépendantes, on calcule R^2 en exprimant cette somme des carrés de la régression comme la proportion de la somme totale des carrés, comme ci-dessous :

$$R^2 = \frac{\text{Somme des carrés de la régression}}{\text{Somme totale des carrés}}$$

$$= \frac{56{,}527}{168{,}694}$$

$$= 0{,}335$$

$$= 0{,}34$$

Pourquoi a-t-on indiqué les dl, la somme moyenne des carrés, *F* et p ? Parce qu'on en aura besoin dans un moment lorsque nous effectuerons des tests de signification pour le R^2.

12.3 Les coefficients de régression standardisés (coefficients bêta)

Les équations de régression décrivent la relation entre une variable dépendante et un groupe de variables indépendantes, et leurs coefficients de régression non standardisés mesurent les effets des variables indépendantes sur les variables dépendantes. Cependant la taille du coefficient de régression dépend des unités de mesure des variables. La disponibilité des radios est par exemple mesurée en nombre de radios pour 100 habitants. Il est clair que le coefficient de régression serait différent si cette variable était mesurée en nombre de radios pour 10 habitants, ou en nombre de radios pour 1 000 habitants.

Ce n'est pas un problème en soi – cela fait partie de la nature des coefficients de régression – mais cela rend cependant la comparaison difficile lorsque les variables sont basées sur des unités de mesures différentes. Par exemple : les moyennes du taux d'urbanisation et du nombre de radios par 100 habitants sont respectivement de 52,4 et 35,9. Leurs écarts-types sont de 25,1 et 39. Comment peut-on comparer ces nombres qui sont basés sur des unités de mesures différentes ? Nous avons ici des pommes et des oranges statistiques ! La sagesse populaire est fondée : on ne peut pas comparer des fruits différents. Pas plus qu'on ne peut comparer des points de pourcentage et des radios par 100 habitants.

Il est clair qu'il nous faut une mesure des effets des variables indépendantes qui tienne compte des différences d'unités de mesure tout en contrôlant les effets des autres variables indépendantes. Cette statistique existe et s'appelle justement ***le coefficient de régression standardisé***, ou de façon plus simple, ***un coefficient bêta***. Les bêtas sont symbolisés par $\beta_{Y1 \bullet 2}$ pour la régression entre Y et X_1 tout en contrôlant l'effet de X_2 (pour des variables additionnelles, on ajoute des

indices après le point (•) selon le besoin). Le coefficient bêta décrit les effets d'une variable indépendante sur la variable dépendante en unités d'écart-type. Plus précisément, le coefficient bêta indique le changement en écarts-types de la variable dépendante pour chaque augmentation d'un écart-type de la variable indépendante, tout en contrôlant les effets des autres variables indépendantes. Voici les formules pour les coefficients bêta avec une variable de contrôle :

$$\beta_{Y1\bullet2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \qquad \text{et} \qquad \beta_{Y1\bullet2} = \frac{r_{Y2} - r_{Y1}r_{21}}{1 - r_{21}^2}$$

Les formules sont plus complexes pour les bêtas d'ordre plus élevé – c'est-à-dire les bêtas avec des variables de contrôle additionnelles. Nous laisserons l'ordinateur les calculer pour nous. Pour notre exemple avec deux variables indépendantes :

$$\beta_{Y1\bullet2} = \frac{-0{,}556 - (-0{,}471)(0{,}617)}{1 - (0{,}617)^2} \quad \text{et} \quad \beta_{Y2\bullet1} = \frac{-0{,}471 - (-0{,}556)(0{,}617)}{1 - (0{,}617)^2}$$

$$= \frac{-0{,}556 + 0{,}291}{1 - 0{,}381} \qquad = \frac{-0{,}471 + 0{,}343}{1 - 0{,}381}$$

$$= \frac{-0{,}265}{-0{,}619} \qquad = \frac{-0{,}128}{-0{,}619}$$

$$= -0{,}428 \qquad = -0{,}207$$

Prenons le coefficient bêta pour le taux d'urbanisation : –0,428. Ce bêta nous indique que la variable dépendante, le taux de fertilité, décroît en moyenne de 0,428 écart-type lorsque la variable indépendante, le taux d'urbanisation, croît d'un écart-type, en contrôlant l'effet du nombre de radios pour 100 habitants. Donc, avec la disponibilité de radios tenue pour constante, une augmentation d'un écart-type dans le taux d'urbanisation entraîne une réduction du taux de fertilité de 0,428 écart-type. On peut trouver ce que représente un écart-type de 0,428 en termes de nombre d'enfants par femme (c'est-à-dire le taux de fertilité) en le multipliant par 1,875, l'écart-type du taux de fertilité : (0,428) (1,875) = 0,80. L'écart-type du taux d'urbanisation est de 25,1. Donc, avec le nombre de radios tenu pour constant, une augmentation de 25 points de pourcentage du taux d'urbanisation est associée avec une baisse d'environ 0,80 enfant par femme.

De même, le coefficient bêta pour la disponibilité des radios de –0,207 signifie qu'en contrôlant les effets du taux d'urbanisation, pour chaque augmentation d'un écart-type du nombre de radios on s'attend à une baisse de 0,207 écart-type du taux de fertilité. Je vous laisse le soin de convertir ce bêta en changement du nombre d'enfants par femme (les écarts-types du taux de fertilité et du nombre de radios sont respectivement de 1,875 et 38,997).

En mesurant les effets des variables en termes d'écarts-types, les bêtas fournissent un moyen utile de comparer les effets relatifs des variables. Nous voyons dans cet exemple que le taux d'urbanisation affecte deux fois plus le taux de fertilité dans les pays les plus peuplés que le nombre de radios par 100 habitants.

En passant, à la différence d'un coefficient de corrélation, un coefficient bêta peut être supérieur à 1 puisqu'un changement d'une unité d'écart-type dans une variable indépendante peut produire un changement de plus d'un écart-type dans la variable dépendante.

12.4 Les tests de signification pour les coefficients de corrélation multiple

Je vous ai déjà donné une liste de plusieurs conditions d'application de la régression multiple : des données d'intervalles ou de proportion, des relations linéaires, des variables indépendantes dont les effets s'additionnent et des variables indépendantes qui ne sont pas trop corrélées entre elles. Nous pouvons faire un test de signification statistique pour des coefficients de corrélation multiple à condition que trois autres conditions soient respectées. Premièrement, nos données proviennent d'un échantillon tiré aléatoirement d'une population. Deuxièmement, les scores de la variable dépendante sont distribués normalement à l'intérieur de chaque valeur de la variable indépendante et de la variable de contrôle et, troisièmement, les variances de la variable dépendante sont égales à l'intérieur de chaque valeur de la variable indépendante et de la variable-contrôle (c'est-à-dire homoscédasticité, comme nous en avons parlé à la section 9.2).

Le test de signification pour le R^2 (donc aussi pour R) utilise la distribution du *F*. Comme dans l'analyse de variance, la valeur du *F* pour R^2 est donnée par la proportion de la somme des carrés moyenne expliquée par la régression sur la somme des carrés moyenne non expliquée par la régression (résidus). On trouve ces sommes des carrés moyennes expliquée et non expliquée en divisant chaque somme des carrés par le nombre approprié de degrés de liberté, qui sont respectivement k et N – k –1. N est comme toujours le nombre de cas

et k est le nombre de variables indépendantes. Dans notre exemple, les degrés de liberté sont alors de k = 2 et N – k – 1 = 46. Pour le R^2 de la régression de la relation entre le taux de fertilité, le taux d'urbanisation et le nombre de radios, on a :

$$\begin{aligned} F &= \frac{\text{Somme des carrés moyenne de la régression}}{\text{Somme des carrés moyenne des résidus}} \\ &= \frac{28{,}263}{2{,}438} \\ &= 11{,}593 \end{aligned}$$

Nous trouvons la probabilité associée à $F(2, 46) = 11{,}593$ dans le tableau 3 de l'appendice ainsi que nous l'avons fait pour les valeurs de *F* dans l'analyse de variance. Bien qu'elle soit basée sur seulement 49 cas, la corrélation multiple de la régression de la relation entre le taux de fertilité, le taux d'urbanisation et le nombre de radios pour 100 habitants est statistiquement significative au seuil 0,001. La question de la généralisation à l'ensemble de la population ne se présente pas ici puisque nous travaillons sur des données de population. Cependant le test de signification nous donne confiance que la relation entre le taux de fertilité et les deux variables indépendantes n'est pas due simplement à l'effet du hasard.

On peut également trouver *F* directement à partir du R^2 avec cette formule :

$$\begin{aligned} F &= \left(\frac{R^2}{1-R^2}\right)\left(\frac{N-k-1}{k}\right) \\ &= \left(\frac{0{,}335}{1-0{,}335}\right)\left(\frac{49-2-1}{2}\right) \\ &= (0{,}504)\,(23) \\ &= 11{,}592 \end{aligned}$$

Comme toujours, rappelez-vous que la signification statistique ne dépend pas seulement de l'intensité de la relation mais aussi du nombre de cas. Même un R^2 très faible sera statistiquement significatif si on a un très grand N. Dans notre exemple, le test de signification est basé sur un petit N (seulement 49 cas) et peut donc être pris au sérieux.

12.5 La régression avec des variables dichotomiques et des variables factices

Les corrélations et régressions multiples peuvent inclure des variables indépendantes dichotomiques. Prenez, par exemple, la variable dichotomique « sexe », avec les hommes codés 0 et les femmes codées 1[1]. Nous avons déjà vu (dans la section 3.5) que la moyenne de la variable « sexe » (0,56 dans le General Social Survey) correspond à la proportion de femmes et (dans la section 4.1) que la variance de « sexe » est la proportion d'hommes multipliée par la proportion de femmes (0,56 x 0,44 = 0,25). Nous avons également appris (dans la section 10.8) que nous pouvons incorporer des variables dichotomiques, même si elles sont nominales (comme par exemple le sexe) dans des équations de régression d'ordre zéro et que l'on peut en tirer des coefficients de corrélation.

En tant que variable indépendante, la variable « sexe » indique l'effet du fait d'être une femme sur la variable dépendante dans une régression d'ordre zéro ou dans une régression multiple. Si l'on utilise les données du General Social Survey, voici l'équation de régression de la relation entre la variable dépendante « prestige lié à la profession » (une mesure de ratio du statut professionnel) et les variables indépendantes « sexe » (X_1) et « niveau d'instruction » (X_2) :

$$Y' = 8{,}44 - 0{,}22\ X_1 + 2{,}56\ X_2$$

lorsque Y' = le score prédit de la variable « prestige lié à la profession »

X_1 = le score de la variable « sexe »

X_2 = le score de la variable « niveau d'instruction »

Nous interprétons les coefficients de régression et de corrélation pour la variable « sexe » de la même façon que nous le faisons pour toute autre variable. Le coefficient de régression « non standardisé » –0,22 indique que les femmes ont un prestige professionnel plus faible que celui des hommes, lorsque que l'on contrôle l'effet du niveau d'instruction. La corrélation multiple au carré de 0,29 est due principalement à l'effet substantiel du niveau d'instruction sur le prestige professionnel. (Incidemment, si l'on avait inversé le codage de la variable « sexe », en codant les femmes 0 et les hommes 1, les signes + ou – des coefficients de corrélation d'ordre zéro et des coefficients de corrélation partielle auraient été inversés, mais la valeur

1. Comme je l'ai souligné dans la section 10.8, les hommes sont codés 1 et les femmes 2 dans le GSS américain. Cela ne change rien à la logique des variables factices, seule change la grandeur des coefficients.

de ces coefficients serait demeurée la même. Les coefficients auraient alors indiqué l'effet d'être un homme sur le prestige lié à la profession.)

L'inclusion de variables indépendantes dichotomiques, même nominales, dans les analyses de régression et de corrélation est assez facile. La méthode est exactement la même que pour les autres variables. Mais les variables nominales qui ont plus de deux valeurs sont, elles, plus compliquées à manipuler. Elles doivent être transformées en ***variables factices*** avant d'être introduites dans le modèle de régression. Une variable factice n'a que deux valeurs, 0 et 1, 0 indiquant l'absence d'un attribut et 1 en indiquant la présence.

Les variables factices viennent en groupes de deux ou plus. Je vais vous montrer comment procéder en prenant pour exemple la variable « religion ». Voici les codes correspondant aux cinq valeurs de cette variable :

0 Protestant
1 Catholique
2 Juif
3 Aucune
4 Autre

Nous créons quatre variables factices qui contiennent l'information concernant la religion du répondant :

Variables factices	Code
Rel.Catho.	1 si Catholique 0 autrement
Rel.Juif	1 si Juif 0 autrement
Rel.Aucune	1 si aucune 0 autrement
Rel.Autre	1 si autre 0 autrement

Les catholiques seront codés 1 pour la variable Rel.Catho. et 0 pour les autres variables factices. Les juifs seront codés 1 seulement pour la variable Rel.Juif et 0 pour les autres variables. Même chose pour ceux qui se disent d'aucune religion ou d'autres religions : ils seront codés 1 pour la variable Rel.Aucune ou Rel.Autre, selon le cas, et 0 pour les autres variables. Remarquez qu'il ne nous faut que *quatre* variables factices pour contenir la totalité de l'information concernant *cinq* préférences religieuses car les répondants qui ont choisi la réponse « protestant » (la cinquième valeur) sont identifiés par le

fait qu'ils sont codés 0 pour toutes les variables factices (il existe une raison technique qui préconise l'utilisation de la valeur modale – dans ce cas « protestant » – comme « valeur de référence » codée 0 pour toute les variables factices). En y réfléchissant un peu, vous comprendrez sans peine pourquoi nous utilisons toujours une variable factice de moins que le nombre de valeurs de la variable originale.

Comme pour les autres variables dichotomiques codées 0 et 1, la moyenne d'une variable factice correspond à la proportion des cas qui sont identifiés par la valeur 1. Pour la variable « religion » que nous avons transformée en variables factices plus haut, la moyenne de la variable factice Rel.Catho. correspond à la proportion de catholiques, la moyenne de Rel.Juif, à la proportion de juifs, etc. Nous pouvons ensuite nous servir de ces quatre variables factices à l'intérieur d'une analyse de corrélation et de régression. C'est-à-dire que l'on peut faire une régression de la relation entre la variable dépendante Y et les variables Rel.Catho., Rel.Juif, Rel.Aucune et Rel.Autre. Les coefficients pour chaque variable factice nous renseignent sur l'effet de cette caractéristique (par exemple, le fait d'être protestant) sur la variable dépendante. Dans la corrélation partielle, ces effets tiennent compte des autres variables de l'analyse.

De plus, si un groupe de variables factices est utilisé sans y ajouter d'autres variables indépendantes, l'intersection (ou la constante) et le coefficient de régression non standardisé décrivent les moyennes de la variable dépendante pour les catégories distinguées par les variables factices. L'intersection est la moyenne de la catégorie codée 0 pour toutes les variables factices. L'intersection ajoutée au coefficient de régression non standardisé d'une variable factice correspond à la moyenne de la variable dépendante pour la catégorie identifiée par cette variable factice. Le tableau 12.1 présente les coefficients pour la régression de la relation entre le niveau d'instruction et les variables factices pour la religion :

Tableau 12.1. Régression de la relation entre le niveau d'instruction et les variables factices pour la religion

Variable	b	Bêta	Moyenne de l'instruction
Rel.Catho.	0,410	0,059	13,510
Rel.Juif	2,268	0,117	15,368
Rel.Autre	1,361	0,100	14,461
Rel.Aucune	0,419	0,046	13,519
Constante	13,100		13,100

$R^2 = 0,02$; $F(3,2887) = 17,309$; $p < 0,001$

Les moyennes ne sont d'habitude pas présentées dans un tableau de résultat de régression tel que celui-ci. Je les ai cependant mises dans la colonne de droite du tableau pour que vous puissiez voir que la moyenne de l'instruction pour chaque groupe religieux est égale à l'intersection additionnée au coefficient de régression non standardisé de la variable factice de ce groupe. La dernière moyenne – 13,100 – est la moyenne pour les protestants, et elle est égale à l'intersection ou constante. Pour ce qui est de la substance, cette régression montre les effets relatifs plus importants des identités religieuses juives et autres sur le niveau d'instruction. Dans l'ensemble, la relation entre la religion et le niveau d'instruction est relativement faible même si elle statistiquement significative. Le R^2 de 0,02 indique que ce groupe de variables factices explique environ 2 % de la variation du nombre d'années d'instruction.

Une fois de plus, nous voyons que les statistiques constituent un réseau d'idées et de techniques interreliées. Les sommes de carrés réapparaissent dans ce chapitre où la régression rejoint l'analyse de variance et où les coefficients de régression et les intersections décrivent les moyennes. La grâce et l'élégance des statistiques tout comme leur rigueur et leur force découlent de ces interconnexions.

12.6 Résumé du chapitre 12

Voici ce que nous avons appris dans ce chapitre :

- La régression multiple approfondit le modèle de la régression bivariée afin d'y inclure des variables indépendantes additionnelles.
- Avec deux variables indépendantes ou plus, le modèle de régression multiple postule que les variables sont des variables d'intervalles ou de proportion, que les relations sont linéaires, que les effets sont additifs (c'est-à-dire sans interaction), et enfin que les variables indépendantes ne sont pas corrélées entre elles.
- Un coefficient de corrélation multiple apprécie l'intensité de la relation entre une variable dépendante et un ensemble de variables indépendantes.
- Le carré du coefficient de corrélation multiple est la proportion de la variation de la variable dépendante qui est expliquée par l'ensemble des variables indépendantes lorsqu'on les considère simultanément.

- Le coefficient bêta décrit le changement en écarts-types de la variable dépendante associé à une augmentation d'un écart-type de la variable indépendante, en contrôlant les effets des autres variables indépendantes.
- Les tests de signification statistique pour les coefficients de corrélation multiples supposent un échantillonnage aléatoire, le caractère normal de la distribution de la variable dépendante à l'intérieur de chacune des valeurs des variables indépendantes, et l'égalité des variances de la variable dépendante à l'intérieur de chacune des valeurs des variables indépendantes.
- Des variables nominales peuvent être intégrées à une analyse de régression et de corrélation à condition qu'elles soient converties en variables factices.
- L'intersection et les coefficients de régression non standardisés des variables factices décrivent les moyennes de la variable dépendante à l'intérieur des catégories distinguées par les variables factices.

Principaux concepts et procédures

Termes et idées

modèle de régression multiple
équation de régression multiple
coefficients de régression non standardisés
interaction statistique
multicolinéarité
coefficient de corrélation multiple
coefficient bêta
variable factice

Symboles

Y'
b_1
b_2
$R^2_{Y \bullet 12...}$
$\beta_{Y1 \bullet 2...}$

Formules

$$Y' = a + b_1X_1 + b_2X_2$$

$$R^2 = \frac{\text{Somme des carrés de la régression}}{\text{Somme totale des carrés}}$$

$$F = \frac{\text{Somme des carrés moyenne de la régression}}{\text{Somme des carrés moyenne des résidus}}$$

$$F = \left(\frac{R^2}{1-R^2}\right)\left(\frac{N-k-1}{k}\right)$$

$$\beta_{Y1\bullet 2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1-r_{12}^2}$$

$$\beta_{Y1\bullet 2} = \frac{r_{Y2} - r_{Y1}r_{21}}{1-r_{21}^2}$$