UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉTUDE DE TESTS DE PERMUTATION EN RÉGRESSION MULTIPLE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

NAOUAL ELFTOUH

FÉVRIER 2008

REMERCIEMENTS

Avant de commencer, je tiens à remercier Madame Sorana FRODA, ma directrice de recherche. Ses conseils judicieux et son encouragement ont été précieux. Sa disponibilité et le temps qu'elle a consacré à répondre à mes questions ont facilité la rédaction de ce mémoire, sans oublier son soutien financier.

Mes sincères remerciements à Monsieur Bertrand Fournier et à Monsieur Fabrice Larribe pour m'avoir aidé à réaliser les simulations.

Enfin, je tiens à remercier mes parents et mon mari qui m'ont grandement soutenu et encouragé constamment à réaliser mes rêves d'accomplir mes études supérieures.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Ce mémoire est centré sur l'étude des coefficients de corrélation partiels en régression linéaire multiple, à travers les tests de permutation. Ces tests sont nécessaires lorsque les suppositions du modèle linéaire ne sont pas verifiées, et l'application des tests classiques est erronnée. On présente les bases théoriques de trois méthodes de la littérature, Manly (1991), Freedman et Lane (1983) et Kennedy (1995), et on fait une étude de simulation afin de les comparer. On ajoute aux comparaisons le test paramétrique, ainsi qu'une méthode qu'on propose. On regarde l'erreur de type I et la puissance de ces tests.

Un dernier volet du mémoire est la présentation des tests de Mantel (1967) et Smoose et al. (1986) qui sont des généralisations de ces méthodes de permutation pour la régression multiple à des matrices de distances. À titre d'exemple, ces différentes techniques de permutation sont appliquées sur des matrices de distances génétiques en relation avec des distances environnementales et des distances géographiques.

Mots clés : Échangeabilité, test de permutation, test de Mantel, test de Mantel partiel, régression multiple, corrélation partielle, résidus.

INTRODUCTION

Lorsque les conditions d'application de l'inférence statistique paramétrique ne sont pas respectées, les résultats des tests paramétriques sont peu fiables. Fisher (1935) et Pitman (1937) ont proposé une nouvelle classe de tests, les tests de permutation. En dépit d'être les plus anciens de tous les tests non paramétriques, ces tests ne se sont pas répandus à cause de l'énorme tâche de calcul qu'ils demandent. Avec l'augmentation de la capacité de calcul des ordinateurs, ces tests sont de plus en plus utilisés dans des situations de plus en plus différentes et des domaines d'application très divers.

Tous les tests de permutation sont dérivés du même principe fondamental. Ce principe est intuitif : sous l'hypothèse nulle appropriée, et en fixant les valeurs échantillonnales, toute permutation des observations de l'échantillon a la même probabilité d'être tirée. On demande que, sous l'hypothèse nulle, les observations soient échangeables (concept un peu plus général que celui de variables indépendantes et de même loi, voir chapitre I).

Ainsi, un test de permutation consiste à une comparaison de la valeur observée de la statistique du test avec les valeurs générées en permutant les données. Les probabilités critiques sont calculées en considérant la loi conditionnelle par rapport au vecteur des statistiques d'ordres observées, et cette loi conditionnelle fournit une distribution empirique par rapport à laquelle la statistique observée du test peut être située.

Les tests de permutation sont parmis les tests les plus puissants quand les suppositions des tests paramétriques traditionnels ne sont pas valides (Good, 1994). Cependant, dans des situations complexes, il n'est pas toujours évident de définir la façon de permuter les données afin d'effectuer ce test. Par exemple, la construction d'un test de permutation pour les coefficients de régression partiels dans la régression multiple est controversée. C'est ce contexte d'application qui est l'objet central de notre étude.

Ainsi, soit un modèle de régression multiple $Y = \beta_1 X + \beta_2 Z + \epsilon$, et supposons qu'on s'intéresse à tester l'effet de la variable X sur Y, en tenant compte de la présence de Z, donc à tester $H_0 : \beta_1 = 0$. Si on voulait effectuer un test à partir de triplets d'observations (x_i, z_i, y_i), $i = 1, \ldots, n$, en permutant les valeurs de Y, cela ne serait pas approprié car les valeurs observées de Y ne sont pas échangeables sous l'hypothèse nulle $H_0 : \beta_1 = 0$. Les différentes méthodes de permutation suggérées dans la littérature (Freedman et Lane, 1983 ; Collins, 1987 ; Oja, 1987 ; Welch, 1990 ; Braak, 1992 ; Kennedy, 1995 ; Manly, 1997) reflètent des différences dans la façon de construire un test de permutation qui est forcément approximatif (voir chapitre 2). Dans ce mémoire, on compare trois méthodes (Kennedy, 1995 ; Manly, 1997 ; Freedman et Lane, 1983) ainsi qu'une nouvelle méthode qu'on a proposée, en terme d'erreur de type I et de puissance, en utilisant des simulations numériques. On présente aussi la généralisation de cette méthodologie à des modèles de régression multiple sur des matrices de distances. Cela comprend, entre autres, le test classique de Mantel (1967).

La structure du mémoire est la suivante. Le premier chapitre présente le principe des tests de permutation, leur condition d'application et leurs propriétés statistiques, ainsi que le test t de Student de permutation et une brève description d'une application de ces tests dans le domaine de la génétique.

Dans le chapitre II, on présente le test de permutation exact pour le coefficient de corrélation partiel en régression linéaire multiple, ainsi que trois méthodes de permutation approximatives de la littérature (Kennedy, 1995 ; Manly, 1997 ; Freedman et Lane, 1983). On suggère aussi notre propre méthode qui utilise une technique de type *bootstrap* pour estimer le paramètre inconnu, estimation nécessaire pour faire un test qui ressemble le plus au test de permutation *exact*.

Le chapitre III compare l'erreur de type I et la puissance empiriques des quatre méthodes présentées dans le chapitre II en utilisant des simulations. On examine l'effet de la taille d'échantillon n, le degré de colinéarité entre les covariables, ρ_{XZ}, la valeur du coefficient de la covariable Z, β_2, et la loi de l'erreur aléatoire, ϵ.

Enfin, le $4^{ème}$ chapitre clôture ce mémoire par une généralisation des tests de permutation à des modéles de régression sur des matrices de distances. On présente la procédure introduite par Mantel (1967), ainsi que les modifications ajoutées par Smoose et al. (1986). Ensuite, on décrit le test de Mantel partiel lorsqu'il y a plus de deux matrices de distances. Finalement, on illustre ces méthodes en appliquant le test de Mantel simple et le test de Mantel partiel sur des données écologiques présentées par Manly (1997).

CHAPITRE I

LES TESTS DE PERMUTATION

Les approches paramétriques standard exigent plusieurs suppositions d'application concernant le plan de l'expérience (échantillonnage aléatoire) et le modèle de population (distribution normale ou homoscédasticité, par exemple). Lorsque les conditions d'application de ces approches ne sont pas respectées, en particulier lorsque la loi des données n'est pas conforme aux exigences du test, les résultats des tests paramétriques sont moins fiables.

Les tests non paramétriques offrent une alternative importante puisqu'ils necessitent moins de suppositions. Une classe importante de tels tests sont les tests de permutation que l'on décrit dans ce premier chapitre. On présente leurs conditions d'application et leurs propriétés statistiques. Afin d'illustrer la méthodologie, on introduit un exemple d'application pour le cas de la comparaison de la moyenne de deux échantillons indépendants. Finalement, on indique l'application de cette méthode dans le domaine de la génétique.

1.1 Principe et propriétés des tests de permutation

L'importance des tests de permutation réside dans leur flexibilité et leur robustesse lorsque les suppositions statistiques des tests paramétriques habituels ne sont pas valides. Ils permettent le choix complet de la statistique du test appropriée au problème en main, et cette liberté de choix permet des milliers d'applications pratiques (voir, par

exemple, Good 1994 et Pesarin 2001).

L'idée derrière ces méthodes est que, au lieu de comparer la valeur d'une statistique de test à une valeur critique correspondant à une distribution théorique de probabilité, on génère la distribution de référence à partir des données mêmes, en recalculant la statistique du test pour chaque permutation des données et en se référant à la loi discrète qui en résulte.

Le principe général de permutation est intuitif. En langage informel il revient à : sous l'hypothèse nulle appropriée, toute permutation des observations de l'échantillon a la même probabilité d'être "tirée" (dans l'ensemble des permutations). En d'autres termes, les différentes permutations possibles de l'échantillon observé sont équiprobables. La propriété d'équiprobabilité est déjà utilisée dans la statistique non paramétrique pour la construction des tests libres de loi (basés sur les rangs). Cette propriété résulte du fait que sous l'hypothèse nulle appropriée, les observations de l'échantillon sont des réalisations de variables aléatoires échangeables (voir définition 1.1).

Pour appliquer les tests de permutation, il faut que la condition nécessaire et suffisante sous l'hypothèse nulle, l'échangeabilité des observations, soit verifiée. Sinon, leur utilisation devient inappropriée et erronée. Afin de détailler ce qu'on vient d'énoncer, on introduit quelques définitions et concepts, et on commence par la définition de variables échangeables (Randles and Wolfe,1979, chapitre I). Dans ce qui suit, on suppose toujours que nos variables sont continues et admettent une densité f.

Définition 1.1

(a) *Soit* Π *l'ensemble des permutations des entiers* $(1, \cdots, n)$. *Un ensemble de variables aléatoires* $X_1, \cdots, X_n$ *est échangeable, si pour toute permutation* $\pi = (\pi_1, \cdots, \pi_n) \in \Pi$ *on a :*

$$(X_1, \cdots, X_n) \quad \underline{d} \quad (X_{\pi_1}, \cdots, X_{\pi_n}), \tag{1.1}$$

où $\underline{d}$ *indique que les deux vecteurs ont la même fonction de répartition (égalité en loi).*

(b) *On dit que la fonction de densité conjointe du vecteur* $\boldsymbol{x} = (x_1, \cdots, x_n)$ *est invariante par rapport aux permutations des éléments du vecteur* $\boldsymbol{x}$ *si :*

$$f(\boldsymbol{x}) = f(x_1, \cdots, x_n) = f(x_{\pi_1}, \cdots, x_{\pi_n}). \tag{1.2}$$

■

L'exemple le plus simple de variables échangeables est le cas où $X_1, \cdots, X_n$ sont des variables indépendantes et identiquement distribuées. Dans ce cas, les variables sont échangeables car :

$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = f(x_{\pi_1}, \cdots, x_{\pi_n}). \tag{1.3}$$

Le problème considéré est celui de tester une hypothèse nulle H_0, concernant la loi de probabilité d'un échantillon, contre une alternative H_1 concernant cette loi. Notons par $f_0(\mathbf{x})$ la densité de $\mathbf{x}$ sous l'hypothèse nulle H_0 et par $f_1(\mathbf{x})$ la densité de $\mathbf{x}$ sous l'alternative H_1. Avant de décrire un test de permutation, on doit d'abord introduire le concept de test statistique.

Définition 1.2

Un test non randomisé Ψ *est une règle de décision (fonction) qui prend les valeurs 0 ou 1,*

$$\Psi(\boldsymbol{x}) = \begin{cases} 1 & \text{on rejette } \mathrm{H}_0; \\ 0 & \text{on accepte } \mathrm{H}_0. \end{cases} \tag{1.4}$$

■

Dans ce cas, faire un test Ψ consiste à déterminer une région de rejet (qui correspond à l'ensemble de $\mathbf{x}$ tel que $\Psi(\mathbf{x}) = 1$) telle que, si l'échantillon observé appartient à cette région, on rejette l'hypothèse nulle H_0. Les critères qui définissent la région de rejet d'un test Ψ sont l'erreur de type I (la probabilité de rejeter H_0 quand elle est vraie) et la puissance (la probabilité de rejeter H_0 chaque fois qu' elle est fausse).

La region de rejet ou critique, C, se définit à l'aide d'une variable aléatoire appelée statistique du test, $U(\mathbf{x})$ et on met

$$\Psi(\mathbf{x}) = 1 \quad si \quad \mathbf{x} \in C\,.$$

On veut toujours contrôler l'erreur de type I pour qu'elle ne dépasse pas un seuil de signification determiné α et maximiser la puissance du test ; ainsi, on met :

$$P(U(\mathbf{x}) \in C) = \int \Psi(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \quad \leq \alpha\,. \tag{1.5}$$

Un tel test s'appele test *conservateur*. Si, en plus, on a que pour toute densité f_0 qui satisfait H_0,

$$\int \Psi(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \quad = \alpha,$$

le test s'appelle *exact*.

Notons par $\mathbf{H}$ la famille de densités *invariantes* par rapport aux permutations, par $\mathbf{S_x}$ l'ensemble qui contient toutes les permutations de $\mathbf{x}$, avec $\mathbf{x}$ fixé et par N le nombre de points de $\mathbf{S_x}$, évidemment on a $N = card\{\mathbf{x}_\pi \in \mathbf{S_x}\} = n!$. Dans ce qui suit on montre que les tests de permutation sont des tests conditionnels et qu'on travaille avec la loi discrète sur l'espace $\mathbf{S_x}$.

Tout d'abord, on commence par définir les statistiques d'ordre et le vecteur des rangs.

Définition 1.3

Soit un vecteur aléatoire $\boldsymbol{X} = (X_1, \cdots, X_n)$.

(a) La statistique d'ordre k, notée $X_{(k)}$, *est égale à la* $k^{ème}$ *plus petite valeur.*

(b) Soit $(X_{(1)}, \cdots, X_{(n)})$ *le vecteur des statistiques d'ordre correspondant à* $\boldsymbol{X} = (X_1, \cdots, X_n)$. *L'observation échantillonale* X_i *a le rang* R_i *parmi* $X_1, \cdots, X_n$ *si* $X_i = X_{(R_i)}$, *à condition que la* R_i*ème statistique d'ordre soit uniquement définie.*

■

Soit $\pi = (\pi_1, \cdots, \pi_n)$ une permutation quelconque dans Π et le vecteur $\mathbf{x}_\pi = (x_{\pi_1}, \cdots, x_{\pi_n})$

correspondant. On présente un théorème important sur lequel se base l'inférence non paramétrique, et en particulier les tests de rangs.

Théorème 1.1 *(Randles and Wolfe,1979, chapitre II, th 2.3.3)*

Soit le vecteur $\boldsymbol{X} = (X_1, \cdots, X_n)$ *de densité* $f_0 \in \boldsymbol{H}$. *Alors le vecteur des rangs* $\boldsymbol{R} = (R_1, \cdots, R_n)$ *et le vecteur des statistiques d'ordre* $\boldsymbol{X}_{(.)} = (X_{(1)}, \cdots, X_{(n)})$ *sont indépendants. De plus,* $\boldsymbol{R}$ *est uniformément distribuée sur* Π, *c'est à dire :*

$$P(\boldsymbol{R} = \pi) = \frac{1}{n!}, \qquad \forall \pi \in \Pi.$$

■

D'après le théorème précédent, l'indépendance entre le vecteur des rangs et le vecteur des statistiques d'ordre nous permet d'écrire :

$$P(\mathbf{X} = \mathbf{x}_\pi \mid \mathbf{X}_{(.)} = \mathbf{x}_{(.)}) = P(\mathbf{R} = \pi \mid \mathbf{X}_{(.)} = \mathbf{x}_{(.)}) = \frac{1}{n!}. \tag{1.6}$$

L'égalité (1.6) s'interprète ainsi : en gardant fixe le vecteur des statistiques d'ordre $\mathbf{x}_{(.)} = (x_{(1)}, \cdots, x_{(n)})$, toutes les $N = n!$ permutations des valeurs du vecteur $\mathbf{x}_{(.)}$ sont équiprobables. On conclut que, sous l'hypothèse H_0 qui suppose l'échangeabilité des variables, la distribution conditionnelle $\{\mathbf{X}|\mathbf{X}_{(.)} = \mathbf{x}_{(.)}\}$ est uniforme sur $\mathbf{S}_\mathbf{x}$. En d'autres mots, sous H_0, on peut considérer que l'échantillon observé $\mathbf{x}$ a la même probabilité d'être choisi que tout autre élement de $\mathbf{S}_\mathbf{x}$, car tous les points de $\mathbf{S}_\mathbf{x}$ sont équiprobables.

La définition formelle d'un test de permutation peut se faire de deux façons : en termes de sa fonction critique, Ψ, ou en termes de statistique de test (via le principe de permutation).

Définition 1.4 *(Hajek 1967, II.2.1)*

Un test de permutation au seuil de signification α a une fonction critique qui s'exprime comme fonction du vecteur des statistiques d'ordre $x_{(\cdot)}$ et des rangs, $\pi = (\pi_1, \pi_2, \ldots \pi_n)$, donc $\Psi(x) = \widetilde{\Psi}(x_{(\cdot)}, \pi_1, \pi_2, \ldots \pi_n)$ et est telle que

$$\frac{\sum_{\pi} \widetilde{\Psi}(x_{(\cdot)}, \pi_1, \pi_2, \ldots \pi_n)}{N} = \alpha.$$

■

Si le test est non randomisé, $\Psi(\mathbf{x})$ ne prend que les valeurs 1 ou 0 et on ne peut pas toujours atteindre α exactement ; alors, la définition précédente s'applique aux valeurs α qui sont de la forme $k/N, k = 1, 2, \ldots, N$.

Remarque 1.1 *(Randles et Wolfe 1979, formule (11.1.4))*

Si on décrit la région de rejet à l'aide d'une statistique de test, $U(\boldsymbol{x})$, cette statistique doit satisfaire le **principe de permutation**, *qui revient au tirage uniforme sous H_0. Si on considère l'ensemble $\mathcal{U} = \{\boldsymbol{u} | \boldsymbol{u} = U(\boldsymbol{x}_\pi), \boldsymbol{x}_\pi \in \boldsymbol{S}_x\}$ on doit avoir que :*

$$P(U(\boldsymbol{X}) = \boldsymbol{u} | \boldsymbol{X}_{(\cdot)} = \boldsymbol{x}_{(\cdot)}) = \frac{1}{N}.$$

■

En d'autres mots, la statistique du test est de loi **conditionnelle** uniforme sur $\mathcal{U}$. Alors, pour un seuil de signification $\alpha = k/N$, $k = 1, \cdots, N$, la région critique va contenir k éléments de $\mathcal{U}$, par exemple les k plus grandes valeurs de U. À la section (1.3) nous décrivons en détail un tel test. Pour conclure, nous signalons les points suivants :

- Les tests de permutation sont des procédures conditionnelles où le conditionnement se fait par rapport au vecteur des statistiques d'ordre et cela donne une loi discrète sur l'espace des permutations $\mathbf{S_x}$.

- Les tests de permutation sont des tests libres de loi parce que la loi de la statistique du test ne depend pas de la densité $f_0(.) \in \mathbf{H}$, qui peut être partiellement ou complètement inconnue.
- Les tests de permutation sont toujours exacts ou conservateurs puisque la probabilité de l'erreur de type I est contrôlée pour tous les échantillons possibles de densité $f_0 \in \mathbf{H}$.

1.2 L'application d'un test de permutation : étapes de calcul et méthode Monte Carlo

Selon Good (1994), pour faire un test de permutation, il faut procéder en quatre étapes, comme suit :

(a) on commence par l'analyse du problème : définir l'hypothèse nulle, l'hypothèse alternative, la loi des observations, et les suppositions du test ;

(b) après, on choisit la statistique de test la plus appropriée pour distinguer les deux hypothèses, et on calcule la statistique pour les observations dont on dispose ;

(c) on recalcule la statistique du test pour toutes les permutations possibles des observations de départ ; notons qu'il faut préciser comment faire les permutations selon chaque problème ;

(d) finalement, on prend la décision : accepter ou rejeter l'hypothèse nulle en utilisant la loi de la statistique comme guide. Cela revient à voir si la statistique observée est extrême ou non par rapport à la loi empirique de la statistique de test générée par les permutations.

Si on considère les étapes (b) et (c) on constate plusieurs problèmes liés aux tests de permutation. Premièrement, la statistique du test sous H_0, particulièrement dans les situations multivariées, peut avoir des formes difficiles à exprimer et à calculer. Un exemple serait le cas où on fait un test de permutation à partir du rapport de vraisemblance. Ensuite, si la taille d'échantillon n'est pas petite, un calcul direct de la loi par énumeration de toutes les permutations devient impraticable à cause du grand

nombre de permutations à faire. Par exemple, même si on compare les espérances de deux échantillons de tailles $m = 5$ et $n = 5$, le nombre de permutations possibles est $\binom{12}{7} = 30240$. Finalement, une approximation asymptotique de la statistique du test n'est pas toujours appropriée, à moins que le nombre d'observations soit suffisamment grand.

Même si, en principe, il est toujours possible de faire un calcul exact en énumérant toutes les permutations possibles, en pratique, on utilise des approches qui sont plus efficaces du point de vue des calculs. La méthode la plus utilisée est la méthode de Monte Carlo. Cette technique consiste à tirer un échantillon aléatoire de taille M, M grand, parmi toutes les permutations possibles et appliquer le test seulement sur cet échantillon, au lieu de considérer la loi complète. En effet, avec un grand échantillon, la distribution sous H_0 peut être bien approximée par la technique de Monte Carlo. Alors, on estime la p-valeur par

$$\widehat{P}_{H_0}(\text{rejetter } H_0) = \frac{k^*(\mathbf{x})}{M}$$

où $k^*(\mathbf{x})$ est le nombre de fois que la statistique du test est plus extrême que la statistique observée parmi les M pemutations retenues dans l'échantillon.

Efron (1993) a demontré que le nombre de valeurs de la statistique de test qui sont extrêmes par rapport à la valeur observée suit une loi binomiale $Bin(M, \alpha)$. Plusieurs auteurs ont signalé (Ernst 2004, Good 1994, Manly 1997) qu'un M égal à quelques milliers est suffisant pour obtenir une estimation précise de la probabilité critique (p-valeur) exacte.

Dans ce qui suit, nous allons illustrer les concepts présentés sur deux exemples.

1.3 Le test de permutation pour deux échantillons indépendants

Afin d'illustrer le principe de permutation, on prend l'exemple de comparaison de deux échantillons indépendants. On a deux groupes d'observations $\{x_i, i = 1, \cdots, n\}$ et $\{y_j, j = 1, \cdots, m\}$ d'effectifs n et m, et on veut tester $H_0 : \mu_X = \mu_Y$ contre l'alter-

native $H_1 : \mu_Y \geq \mu_X$, par exemple. On choisit comme statistique $U(\underset{\sim}{X}, \underset{\sim}{Y}) = \overline{Y} - \overline{X}$.

Pour faire un test de permutation, il faut générer la loi conditionnelle de la statistique U sous H_0. D'abord, on calcule la valeur observée $u_{obs} = \overline{y}_{obs} - \overline{x}_{obs}$. On prend ensuite tous les $(m+n)$ éléments des vecteurs observés $(x_1, \cdots, x_n)$ et $(y_1, \cdots, y_m)$ et on crée le vecteur $(z_{(1)}, \cdots, z_{(m+n)})$ des statistiques d'ordre des $(m+n)$ observations mises ensembles. Après, on les mélange (permute aléatoirement) et on tire au hasard deux nouveaux groupes avec m et n éléments respectivement. Le nombre de valeurs distinctes de U est égal à $\binom{m+n}{m}$ car si on permute les observations à l'interieur de chaque groupe la valeur de U reste la même.

Si les tailles des échantillons sont petites, on peut faire toutes les permutations possibles. Sinon, on obtient un bon résultat en effectuant seulement un certain nombre M de permutations, par exemple $M = 4999$, choisies au hasard et sans remise parmi toutes celles qui sont possibles.

Après avoir généré les valeurs de la statistique U, on place en ordre croissant les valeurs obtenues $(U_{(1)}, \cdots, U_{(M)})$. Pour un seuil de signification α, le test rejette H_0 si $U_{obs} \geq U_{(k)}$ où $U_{(k)}$ est le quantile qui correspond à $k = [M(1-\alpha)]$, la partie entière de $M(1-\alpha)$. Cela veut dire qu'on rejette si la valeur observée U_{obs} est suffisamment extrême qu'elle ne peut être due seulement au hasard.

On reprend ici l'exemple numérique présenté par Randles et Wolfe (1979), chapitre XI. Les valeurs observées sont $\mathbf{x} = (4.3, 6.0, 3.6)$ et $\mathbf{y} = (7.4, 5.5, 6.2)$. Le vecteur des statistiques d'ordre pour l'échantillon combiné est $\mathbf{z}_{(.)} = (3.6, 4.3, 5.5, 6.0, 6.2, 7.4)$. Le nombre de permutations possibles est $M = \binom{6}{3} = 20$. Le tableau (1.1) présente les différentes valeurs $\mathbf{x}$ et $\mathbf{y}$ possibles étant donné $\mathbf{z}_{(.)} = (3.6, 4.3, 5.5, 6.0, 6.2, 7.4)$, ainsi que la loi discrète de la statistique U. La valeur observée est $t_{obs} = \overline{y}_{obs} - \overline{x}_{obs} = 6.4 - 4.63 = 1.7667$ et, pour chaque valeur u dans la colonne 8, on a

$$P_{H_0}\left(U(\mathbf{X}, \mathbf{Y}) = u \mid Z_{(.)} = (3.6, 4.3, 5.5, 6.0, 6.2, 7.4)\right) = \frac{1}{20}.$$

En effet, en vertu du théorème 1.1, sous H_0 toutes les répartitions des valeurs du vecteur

$\mathbf{z}_{(.)} = (3.6, 4.3, 5.5, 6.0, 6.2, 7.4)$ entre les $\mathbf{x}$ et les $\mathbf{y}$ sont de même probabilité 1/20. Pour un seuil de signification $\alpha = 0.05$ on a $k = M(1-\alpha) = 20(1-0.05) = 19$. Comme la valeur observée $t_{obs} = 1.7667$ est supérieure à la valeur critique $t_{(19)} = 1.733$, on peut rejetter l'hypothèse nulle.

Tableau 1.1 Les permutations possibles de deux groupes de tailles $m = 3$ et $n = 3$

No	X_1	X_2	X_3	Y_1	Y_2	Y_3	$U(\mathbf{x},\mathbf{y})$	Rang
1	3.6	4.3	5.5	6.0	6.2	7.4	2.067	20
2	3.6	4.3	6.0	5.5	6.2	7.4	1.733	19
3	3.6	4.3	6.2	5.5	6.0	7.4	1.600	18
4	3.6	4.3	7.4	5.5	6.0	6.2	0.800	17
5	3.6	5.5	6.0	4.3	6.2	7.4	0.933	16
6	3.6	5.5	6.2	4.3	6.0	7.4	0.800	15
7	3.6	5.5	7.4	4.3	6.0	6.2	0.000	14
8	3.6	6.0	6.2	4.3	5.5	7.4	0.467	13
9	3.6	6.0	7.4	4.3	5.5	6.2	-0.333	12
10	3.6	6.2	7.4	4.3	5.5	6.0	-0.333	11
11	4.3	5.5	6.0	3.6	6.2	7.4	-0.467	10
12	4.3	5.5	6.2	3.6	6.0	7.4	0.333	9
13	4.3	5.5	7.4	3.6	6.0	6.2	-0.467	8
14	4.3	6.0	6.2	3.6	5.5	7.4	0.000	7
15	4.3	6.0	7.4	3.6	5.5	6.2	-0.800	6
16	4.3	6.2	7.4	3.6	5.5	6.0	-0.933	5
17	5.5	6.0	6.2	3.6	4.3	7.4	-0.800	4
18	5.5	6.0	7.4	3.6	4.3	6.2	-1.600	3
19	5.5	6.2	7.4	3.6	4.3	6.0	-1.733	2
20	6.0	6.2	7.4	3.6	4.3	5.5	-2.067	1

1.4 Intérêt pratique des tests de permutation dans la détection d'un QTL en génétique

Pour illustrer l'intérêt pratique de ces tests, nous présentons une application dans le domaine de la génétique (Doerge et Churchill, 1996), mais sans rentrer dans les détails techniques. On commence par définir quelques termes génétiques nécessaires pour comprendre cet exemple. On a :

- ADN : code d'information génétique. les gènes composés de séquences d'ADN, se présentent en paires ;
- Locus : emplacement précis d'un gène sur un chromosome ;
- Marqueur : séquence d'ADN repérable spécifiquement ;
- Allèle : une des différentes formes que peut prendre un même gène ; les allèles occupent la même position (locus) sur les chromosomes appariés ;
- QTL : un locus dont les allèles déterminent la valeur d'un caractère quantitatif, par exemple, le taux de cholestérol.

Pour un marqueur donné, on veut savoir s'il y a un gène à son voisinage responsable de la valeur du caractère quantitatif. Les données observées se composent d'une partie génétique (carte génétique des individus observés, un codage numérique de son ADN) et d'une partie phénotypique qui représente la valeur du caractère quantitatif d'intérêt, par exemple, le taux de cholesterol dans le sang. On a n individus et (y_i, M_i), $i = 1, \cdots, n$, où y_i est la valeur du caractère quantitatif chez l'individu i, et M_i représente le marqueur de l'individu i.

À un point précis du génome (marqueur M) on veut tester l'hypothèse nulle suivante (en langage informel), H_0 : *QTL absent ou présent mais non relié au marqueur*, contre une alternative H_1 : *QTL relié au marqueur*. On suppose que, sous H_0, en réassignant aléatoirement chaque valeur de trait y_i (de l'individu i) à un autre individu qui a la carte génétique j, la loi de Y ne change pas. Pour simplifier l'exemple on considère que la valeur du caractère Y est définie par une composante génétique et une composante aléatoire.

Sans rentrer dans les détails techniques liés à la génétique, le test de détection d'un QTL revient à faire un test de $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, où Δ représente l'effet de la présence du QTL au marqueur en question. Les méthodes de permutation offrent un choix multiple de la statistique de test U : on peut par exemple utiliser le «Lod score» qui est le logarithme du rapport de vraissemblance, la statistique t de Student ou celle de Fisher (les auteurs ont utilisé la statistique t de Student pour cette étude).

L'idée est que, s'il n'y a pas d'effet du QTL lié au marqueur M, on peut mélanger aléatoirement les valeurs du caractère quantitatif y_i à travers les individus sans que la valeur de la statistique de test U devienne extrême.

Pour trouver la loi de la statistique de test U, on numérote les individus de 1 à n, et on calcule la statistique observée u_{obs}. Les valeurs de trait y_i sont permutées (tout en gardant les cartes génétiques fixes) en faisant une permutation $\pi = (\pi_1, \cdots, \pi_n)$ et en assignant la $i^{ème}$ valeur du caractère après permutation, y_{π_i}, à l'individu qui a l'indice i. On répète cette procédure un grand nombre de fois ($M = 4999$ est suffisant) pour trouver les valeurs de la statistique du test $U_{(1)}, U_{(2)}, \cdots, U_{(M)}$, et on cherche le $(1-\alpha)^{ème}$ quantile $U_{[(1-\alpha)M]}$.
La décision du test se fait en comparant u_{obs} avec $U_{[(1-\alpha)M]}$ et en rejetant l'hypothèse nulle si $u_{obs} > U_{[(1-\alpha)M]}$. Si on rejette, on conclut que le marqueur est relié au caractère quantitatif d'intérêt. On refait le test de permutation pour tous les marqueurs afin de trouver les QTL qui contribuent à la détermination de la valeur du caractère quantitatif.

CHAPITRE II

LES TESTS DE PERMUTATION ET LA RÉGRESSION LINÉAIRE MULTIPLE

Dans ce chapitre, on va présenter le test de permutation exact pour le coefficient de corrélation (coefficient de régression) dans la régression linéaire simple, ainsi que trois méthodes de permutation approximatives pour le test du coefficient de régression partiel. On suggère aussi notre propre méthode qui utilise la technique du «bootstrap» pour estimer le paramètre inconnu nécessaire pour effectuer le test exact.

Pour introduire le problème, on va commencer par le cas de la régression linéaire simple, qui est suivi de la régression multiple.

2.1. La régression linéaire simple

La situation de régression linéaire simple se produit lorsque deux variables X et Y sont mesurées sur n sujets. Les données se composent de n couples d'observations $(x_1, y_1), \cdots, (x_n, y_n)$ qui sont des réalisations de couples de variables aléatoires (X_i, Y_i) où $i = 1, \cdots, n$. Pour des fins de simplicité, et sans perte de généralité, on considère que les variables X_i et Y_i ont des moyennes nulles, $\mu_X = \mu_Y = 0$. Ainsi, on considère un modèle de régression linéaire simple décrit par l'équation suivante :

$$Y_i = \beta X_i + \epsilon_i, i = 1, \cdots, n. \tag{2.1}$$

Les suppositions de ce modèle sont :

- les erreurs aléatoires ϵ_i, $i = 1, \cdots, n$ sont des variables indépendantes et identiquement distribuées de loi normale $N(0, \sigma_\epsilon^2)$;
- les erreurs ϵ_i sont indépendantes des variables explicatives $X_i, i = 1, \cdots, n$.

Par contre, l'inférence statistique habituelle suppose que les variables X_i sont connues sans erreur. Cela veut dire qu'en régression simple on traite les valeurs de la variable X comme étant fixes (ou, autrement dit, on conditionne par rapport à X).

L'estimateur de la méthode des moindres carrés du coefficient de régression β est décrit par l'équation (2.2)

$$\hat{\beta} = b_{yx} = \frac{S_{yx}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{2.2}$$

où (x_i, y_i), $i = 1, \cdots, n$ sont les valeurs observées. Les résidus estimés sont la différence entre la valeur de Y observée et estimée. Soit :$\widehat{\epsilon_i} = y_i - \widehat{y_i} = y_i - b_{yx}\, x_i$.

Souvent, on a intérêt à savoir si le coefficient de régression β est significativement différent de zéro. Dans ce cas, on fait le test de l'hypothèse nulle $H_0 : \beta = 0$ contre l'alternative $H_1 : \beta \neq 0$.

On sait que le coefficient de corrélation entre les variables X et Y est égal à

$$\rho_{XY} = \frac{\mathrm{cov}(X, Y)}{\sigma_X\, \sigma_Y} = \beta\, \frac{\sigma_X}{\sigma_Y}.$$

Ainsi, sous le modèle (2.1), l'hypothèse $H_0 : \beta = 0$ est équivalente à l'hypothèse $\widetilde{H}_0 : \rho_{XY} = 0$.

Dans la littérature on considère aussi les suppositions un peu plus générales, que les erreurs aléatoires ϵ_i sont des variables identiquement distribuées de loi symétrique par rapport à 0 et telles que $\mathrm{cov}(\epsilon_i, X_i) = 0$, $i = 1, \cdots, n$ et $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$.

2.1.1 Le test paramétrique

Si les suppositions du modèle linéaire (2.1) sont verifiées, et on conditionne par rapport à X alors le test paramétrique usuel est le test t de Student, de statistique observée

définie par l'équation suivante

$$T_{\text{obs}} = \frac{b_{yx}}{Se(b_{yx})} \sim T_{n-2}, \tag{2.3}$$

où T_{n-2} dénote une variable aléatoire t de Student avec $(n-2)$ degrés de liberté et $Se(b_{yx})$ est l'estimateur de l'écart type de b_{yx}. On rejette l'hypothèse nulle si $|t_{\text{obs}}| \geq t_{\alpha,n-2}$, où $t_{\alpha,n-2}$ représente la valeur critique correspondant à un seuil de signification α fixé d'avance.

2.1.2 Le test de permutation

Quand les suppositions du modèle de régression linéaire ne sont pas verifiées, on ne peut pas utiliser le test paramétrique. Dans ce cas, on peut effectuer un test de permutation qui nécessite seulement l'échangeabilité des observations (X_i, Y_i)(voir la définition 1.1).

En effet, sous $H_0 : \beta = 0$, le modèle devient $Y_i = \epsilon_i$ et la supposition nécessaire et suffisante pour faire le test de permutation (l'échangeabilité des Y_i) est verifiée. Donc, sous H_0, les variables X et Y ne sont pas reliées linéairement et tout appariement des valeurs de X et Y est équiprobable. Ainsi, en permutant les valeurs y_i, on obtient $n!$ permutations possibles, en gardant les x_i fixes des n couples $(x_1, y_1) \cdots (x_n, y_n)$, qui sont toutes équiprobables.

Notons d'abord que, dans ce qui suit, on va indiquer par un indice supérieur π que la variable Y (ou le résidu) a été permuté et par un indice inférieur π une valeur Y qui a été recréée à partir d'un résidu permuté.

D'aprés l'équation (2.2), on remarque que les tests basés sur les statistiques S_{yx}, b_{yx} et $r_{yx} = S_{yx}/\sqrt{S_{yy}S_{xx}}$ sont équivalents puisque S_{yy} et S_{xx} restent fixes à travers les permutations. Donc, le test de permutation donnera le même résultat. On va prendre la statistique r_{yx}^2 parce qu'elle est facile à interpreter pour faire le test bilatéral, en utilisant M permutations (méthode Monte Carlo).

Soit (x_i, y_i), $i = 1, \cdots, n$, et soit un sous ensemble $\{\pi^{(1)}, \cdots, \pi^{(M)}\}$ tiré aléatoirement de l'ensemble des permutations possibles, Π, $(M < n!)$. Fixons la permutation $\pi^{(j)}$, $j =$

$1, \cdots, M$. Si on remplace les couples (x_i, y_i), $i = 1, \cdots, n$ par des couples où les valeurs y_i ont été pérmutés selon $\pi^{(j)}$ on obtient une autre valeur r_j^2. Ainsi $\{r_1^2, \cdots, r_M^2\}$ sont les valeurs de la statistique du test pour ce sous-ensemble de permutations et $\{r_{(1)}^2, \cdots, r_{(M)}^2\}$ est l'ensemble en ordre croissant de ces valeurs de r^2.

Alors, le test de permutation est le suivant : on rejette H_0 si $r_{\text{obs}}^2 \geq r_{(k)}^2$, où $r_{(k)}^2$ est le quantile qui correspond à $k = [M(1-\alpha)]$, la partie entière de $M(1-\alpha)$.

Si, dans la littérature, les chercheurs sont d'accord sur la procédure de permutation à utiliser dans le cas de la régression simple (Manly 1991 et Edgington 1987), ce n'est pas le cas lorsqu'il y a plusieurs variables explicatives. Pour simplifier, dans la section suivante, on va discuter du cas de la régression multiple avec deux variables explicatives univariées (X, Z). Les résultats peuvent être généralisés au cas de deux vecteurs de variables $(\mathbf{X}, \mathbf{Z})$ avec $\mathbf{X} = (X_1, \cdots, X_p)$ et $\mathbf{Y} = (Y_1, \cdots, Y_q)$.

2.2 La régression multiple : Tests de ce modèle

Supposons qu'on a mesuré trois variables (X, Y, Z) pour n individus et on a observé (x_i, y_i, z_i), $i = 1, \cdots, n$. Pour la simplicité et sans perte de généralité, on suppose que ces variables ont des espérances nulles $\mu_X = \mu_Y = \mu_Z = 0$. Le modèle de régression linéaire multiple de Y sur X et Z est défini alors par l'équation suivante

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \epsilon_i, \tag{2.4}$$

où Y_i est une variable aléatoire à expliquer, X_i et Z_i sont des variables explicatives, β_1 et β_2 sont les coefficients de régression partiels et ϵ_i est l'erreur aléatoire. Les conditions usuelles pour l'utilisation du modèle de régression linéaire multiple sont

- les erreurs aléatoires ϵ_i, $i = 1, \cdots, n$ sont des variables indépendantes et identiquement distribuées de loi normale $N(0, \sigma_\epsilon^2)$;
- les erreurs ϵ_i sont indépendantes des vecteurs de variables explicatives (X_i, Z_i), $i = 1, \cdots, n$.

Par contre, en inférence, on conditionne par rapport à X et Z, et on suppose X_i, Z_i fixes, $i = 1, \cdots, n$. Notons qu'on peut considérer les suppositions un peu plus générales suivantes

- les erreurs aléatoires ϵ_i sont des variables identiquement distribuées de loi symétrique par rapport à 0 et telles que $cov(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$;
- $cov(X_i, \epsilon_i) = cov(Z_i, \epsilon_i) = 0, i = 1, \cdots, n$.

On entend par test de signification du modèle de régression multiple, le test de l'hypothèse

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{contre} \quad H_1 : \text{au moins } \beta_1 \text{ ou } \beta_2 \neq 0.$$

En d'autres mots, on veut tester l'hypothèse que la variable de réponse Y est indépendante des variables X et Z contre l'hypothèse que la variable Y est (linéairement) liée à au moins une des deux variables X et Z.

Le test paramétrique usuel est basé sur la statistique de Fisher. L'expression la plus commode de cette statistique F est basée sur le coefficient de détermination R^2 qui donne la proportion de la variation totale de Y expliquée par le modèle. Alors on a

$$F = \frac{(n-3)R^2}{2(1-R^2)} \sim F_{2,n-3}, \tag{2.5}$$

où

$$R^2 = \sum(\hat{Y}_i - \overline{Y})^2 / \sum(Y_i - \overline{Y})^2.$$

avec $\hat{Y}_i = b_1 X_i + b_2 Z_i$ et $F_{2,n-3}$ dénote une variable de Fisher avec 2 et $n-3$ degrés de liberté. Ainsi, pour un seuil de signification α, on compare la statistique observée avec la valeur critique correspondante, $F_{\alpha,2,n-3}$.

Afin de construire le test de permutation correspondant, on peut utiliser la statistique R^2 ou F. D'abord, on doit connaître quelles variables sont échangeables sous l'hypothése nulle. Sous H_0, le modèle devient $Y_i = \epsilon_i$. Ainsi, la supposition des erreurs ϵ_i indépendantes et identiquements distribuées peut être allégée pour une supposition d'erreurs échangeables. Dans ce cas, on peut faire un test de permutation exact en permutant les valeurs de la variable Y.

Sous l'hypothèse nulle, Y n'a aucune relation avec X et Z pris ensemble. Cela signifie que, pour des valeurs $X_i = x_i$ et $Z_i = z_i$ fixes, on peut obtenir n'importe quelle valeur de Y_i. En d'autres termes, si on prend des permutations $\pi \in \Pi$, tous les triplets (y_i^π, x_i, z_i) sont équiprobables (conditionnellement aux statistiques d'ordre de $y_i, i = 1, \cdots, n$).

Aprés avoir généré et placé en ordre croissant les M valeurs de la statistique, le test rejette H_0 si $R^2_{\text{obs}} \geq R^2_{(k)}$, où où $R^2_{(k)}$ est le quantile qui correspond à $k = \lfloor M(1-\alpha) \rfloor$, la partie entière de $M(1-\alpha)$.

2.3 Coefficient de régression partiel

Les chercheurs sont souvent intéressés à des hypothèses spécifiques concernant les coefficients de régression partiels, par exemple $H_0 : \beta_1 = 0$. En d'autres termes, on veut savoir si la variable X explique une partie de la variabilité de Y en tenant compte de l'effet de la variable concomitante Z, qui peut être reliée à X.

Si les suppositions du modèle sont verifiées et on conditionne par rapport à X et Z, la statistique du test paramétrique familier prend la forme habituelle décrite par l'équation (2.6)

$$t = \frac{b_1}{Se_{(b_1)}} \sim T_{n-3}, \tag{2.6}$$

où b_1 est l'estimateur des moindres carrés du coefficient β_1, $Se_{(b_1)}$ est l'estimateur de son écart-type et T_{n-3} représente la loi t de Student avec $n-3$ degrés de liberté.

Lorsque les conditions du modéle de régression linéaire multiple ne sont pas respectées, les tests paramétriques ne sont pas fiables. Pour remédier à cette situation, plusieurs chercheurs ont proposé des techniques de permutation approximatives nécessitant peu de suppositions (Welch 1990, Braak 1992, Oja 1987, Brown et Martiz 1982, Freedman et Lane 1983, Kennedy 1995, Manly 1997).

Ici on s'intéresse à présenter des méthodes de permutation qui permettent une relation potentielle entre les variables X et Z (colinéarité). Ces méthodes sont :

1. permutation des résidus sous le modèle réduit proposée par Freedman et Lane (1983) ;
2. permutation des résidus sous le modèle réduit proposée par Kennedy (1996) ;
3. permutation de la variable dépendante Y_i proposée par Manly (1997).

Une autre méthode de la litterature qu'on ne décrit pas ici (car trop différente) est celle proposée par Braak (1992). Afin d'introduire ces trois méthodes (qui sont des tests de permutation approximatifs), on présente un test conceptuel de permutation exact, qui suppose que la relation entre la variable dépendante Y et la variable concomitante Z est connue. À travers ce test conceptuel on peut juger de la qualité des méthodes de permutation appliquées en pratique. Ce développement suit celui présenté dans Anderson et Robinson (2001).

2.3.1 Un test de permutation exact

Dans cette section, nous présentons des idées de tests de permutation *exacts* (dans le sens du chapitre 1) de l'hypothèse nulle $H_0 : \beta_1 = 0$ dans le modèle (2.4). Un tel type de test est conceptuel, car il suppose la connaissance parfaite des paramètres du modèle. On le présente afin de justifier les tests utilisés en pratique.

Considérons le coefficient de corrélation partiel défini en (2.15). Alors, on peut noter que l'hypothèse $H_0 : \beta_1 = 0$ revient à l'hypothèse $\rho_{XY.Z} = 0$ (voir section 2.4.2), et que le carré de ce coefficient s'écrit comme,

$$\rho^2_{XY.Z} = \frac{(\rho_{XY} - \rho_{XZ}\rho_{YZ})^2}{(1-\rho^2_{XZ})(1-\rho^2_{YZ})} = \rho^2_{\epsilon_{XZ}\epsilon_{YZ}}, \tag{2.7}$$

où

$$Y = \lambda_Y Z + \epsilon_{YZ}, \quad X = \lambda_X Z + \epsilon_{XZ}.$$

En d'autres mots, $\lambda_Y Z, \lambda_X Z$ sont les meilleurs approximants linéaires de Y, respectivement X, en fonction de Z (voir section 2.4.2).

L'estimateur de $\rho^2_{XY.Z}$ est $r^2_{XY.Z}$ et est égal à :

$$r^2_{XY.Z} = \frac{(r_{XY} - r_{XZ}r_{YZ})^2}{(1 - r^2_{XZ})(1 - r^2_{YZ})} = r^2_{R_{X|Z}R_{Y|Z}} = \frac{(\sum R_{Y|Z}R_{X|Z})^2}{\sum R^2_{Y|Z} \sum R^2_{X|Z}}, \tag{2.8}$$

où $R_{Y|Z}$ est le résidu de la régression linéaire de Y sur Z et $R_{X|Z}$ est le résidu de la régression linéaire de X sur Z.

Pour effectuer un test *exact* on devrait pouvoir recalculer la statistique $r^2_{XY.Z}$ sous des permutations qui font que toutes les valeurs de $r^2_{XY.Z}$ obtenues sont de même probabilité (principe de **permutation** énoncé au chapitre 1). Supposons qu'on observe (x_i, z_i, y_i), $i = 1, \ldots, n$, et qu'on conditionne par rapport à (x_i, z_i). Ainsi, on fixe les couples (x_i, z_i) et on note que, sous $H_0 : \beta_1 = 0$, les différences

$$Y_i - \beta_2 z_i = Y_i - \lambda_Y z_i = \epsilon_{YZ}$$

sont des variables i.i.d. Donc, si on connaissait λ_Y, on pourrait considérer les triplets $(x_i, z_i, y_i - \lambda_Y z_i)$ et permuter $\epsilon_{YZ} = R_{Y|Z} = y_i - \lambda_Y z_i$ tout en laissant (x_i, z_i) fixes. Pour la simplicité, on omet l'indice i dans $R_{Y|Z}$ et ϵ_{YZ}. Sous H_0, $\epsilon_{YZ} = y_i - \lambda_Y z_i$ sont des réalisations de variables i.i.d. et, par conséquent, les valeurs permutées $\epsilon^\pi_{YZ} = R^\pi_{Y|Z}$ de $\epsilon_{YZ} = R_{Y|Z}$ sont équiprobables, une fois qu'on connaît les statistiques d'ordre de $\epsilon_{YZ} = y_i - \lambda_Y z_i$. Cela revient à dire que tous les $n!$ triplets $(X, Z, R^\pi_{Y|Z})$ avec (X, Z) fixés et $R_{Y|Z}$ permutés sont équiprobables. Dans cette démarche, les valeurs $(x_i, z_i), i = 1, \ldots, n$ restent fixes, donc les résidus $R_{X|Z} = x_i - \hat{\lambda}_X z_i$ ne changent pas. On pourrait même supposer $\hat{\lambda}_X = \lambda_X$ si on considérait seulement la loi discrète, uniforme sur les n couples (x_i, z_i), ce qui donnerait exactement $R_{X|Z} = \epsilon_{XZ}$. En pratique, l'idée de permuter de cette façon a conduit au test de Kennedy (1996), qu'on présente à la section 2.3.4.

Par contre, ici on met l'emphase sur un autre type de test de permutation *exact* dans le sens du chapitre I, qui est celui introduit dans Anderson et Robinson (2001) afin de justifier et comparer les méthodes de Freedman et Lane (1983) et Manly (1997). Pour effectuer le test on passe par les étapes suivantes (l'indice E veut dire exact) :

(1°) on permute les résidus $R_{Y|Z}$ et, à partir des résidus permutés, $R^\pi_{Y|Z}$, on crée de nouvelles observations, $Y_{\pi(E)} = \lambda_Y Z + R^\pi_{Y|Z}$;

(2°) on recalcule les résidus $R_{Y_{\pi(E)}|Z}$ qui correspondent à la régression du vecteur $Y_{\pi(E)}$ sur le vecteur Z ; ces résidus s'écrivent comme $R_{Y_{\pi(E)}|Z} = Y_{\pi(E)} - \hat{\lambda}_{\pi(E)} Z$ où $\hat{\lambda}_{\pi(E)} = \sum Y_{\pi(E)} Z / \sum Z^2$ est l'estimateur de λ_Y pour la permutation π ;

(3°) on recalcule $r^2_{XY.Z}$ en remplaçant dans (2.8) $R_{Y|Z}$ par $R_{Y_{\pi(E)}|Z}$, ce qui donne la statistique de test

$$r^2_E = \frac{(\sum R_{Y_{\pi(E)}|Z} R_{X|Z})^2}{\sum (R_{Y_{\pi(E)}|Z})^2 \sum R^2_{X|Z}} . \tag{2.9}$$

Cette façon de construire un nouveau vecteur d'observations, $Y_{\pi(E)}$, fait en sorte que, si on fixe Z et on conditionne par rapport aux statistiques d'ordre des résidus $R^{\pi}_{Y|Z}$, tout vecteur $Y_{\pi(E)}$ a la même chance d'être tiré. Notons qu'il y a en tout (au plus) $n!$ vecteurs $Y_{\pi(E)}$ différents et, par conséquent, $n!$ valeurs différentes de r^2_E. Par la construction du vecteur $Y_{\pi(E)}$ et le fait que la loi conditionnelle des résidus observés $R^{\pi}_{Y|Z}$ est uniforme (sous H_0), on obtient indirectement que les valeurs de r^2_E sont, elles-aussi, équiprobables (sur l'espace des $n!$ valeurs différentes de r^2_E).

Notons que, dans le calcul de $r^2_{XY.Z}$, à chaque permutation on refait la régression et on obtient les résidus avec λ_Y estimé. Cela se justifie (Anderson et Robinson, 2001) par la nécessité de reproduire chaque fois la démarche effectuée pour obtenir la statistique observée, r^2_{obs}. L'avantage de cette méthode par rapport à celle où on permute seulement les résidus de départ et on travaille avec $R^{\pi}_{Y|Z}$ ressort dans les propriétés des versions pratiques de ces tests de permutation *exacts* (test de Kennedy, 1996 versus Freedman et Lane, 1983).

Cependant, λ_Y est inconnu. Il est donc impossible de faire le test exact. Plusieurs chercheurs ont suggéré des méthodes de permutation approximatives pour remedier à cette situation (Freedman et Lane, 1983 ; Kennedy, 1996 ; Manly, 1997). Dans ce qui suit, on va décrir ces procédures qui sont utilisées en pratique.

2.3.2 L'approche de Freedman et Lane

Freedman et Lane (1983) considèrent que leur méthode est une approche non stochastique en se référant à la probabilité critique (la proportion des valeurs de la statistique qui sont supérieures à la statistique observée) comme étant une statistique descriptive et non une probabilité. Cependant, on peut justifier leur méthode à partir du deuxième test exact décrit à la section précédente.

L'idée du test de Freedman et Lane (1983) est intuitive. Puisque λ_Y est inconnu, on peut l'estimer en régressant Y sur Z. Les résidus sont approximativement égaux aux vraies erreurs dans le modèle de la régression, et sont échangeables sous l'hypothèse nulle. Ainsi, la distribution générée par permutation approxime la distribution qu'on devrait avoir si on avait utilisé les vraies erreurs (Anderson et Robinson, 2001).

Ensuite, pour chacune des $n!$ permutations possibles, la valeur de la corrélation partielle au carré, notée r_F^2, est calculée de la même façon que r_E^2 où λ_Y est remplacée son estimateur $\hat{\lambda}_Y$ dans la partie (1°) du test exact.

Notons que X et Z ne sont pas permutées et restent dans l'ordre original. Cela assure que la relation entre X et Z reste intacte et n'affecte pas le test (les résidus $R_{X|Z}$ restent constants à travers l'ensemble des permutations). Les différentes étapes de l'approche proposée par Freedman et Lane(1983) sont(l'indice F identifie Freedman et Lane) :

Etape I : Calcul de la statistique observée :

(1°) chacune des variables Y et X est régressée sur la variable Z ;

(2°) on calcule $\hat{\lambda}_Y$ et $\hat{\lambda}_X$;

(3°) on obtient les résidus $R_{Y|Z} = Y - \hat{\lambda}_Y Z$ et $R_{X|Z} = X - \hat{\lambda}_X Z$;

(4°) on calcule la statistique du test observée r^2_{obs} à partir de l'équation (2.8) ;

Etape II : Algorithme de permutation :

(1°) on permute aléatoirement les résidus $R_{Y|Z}$ pour trouver le vecteur $R^{\pi}_{Y|Z}$. On recalcule les valeurs $Y_{\pi(F)} = \hat{\lambda}_Y Z + R^{\pi}_{Y|Z}$. L'ancien couple (Y, Z) est remplacé par $(Y_{\pi(F)}, Z)$;

(2°) on régresse $Y_{\pi(F)}$ sur Z, ce qui donne un nouvel estimateur $\hat{\lambda}_{\pi(F)}$ de λ_Y, et après on calcule les résidus $R_{Y_{\pi(F)}|Z} = Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z$;

(3°) on calcule la statistique du test r_F^2

$$r_F^2 = \frac{(\sum R_{Y_{\pi(F)}|Z} R_{X|Z})^2}{\sum (R_{Y_{\pi(F)}|Z})^2 \sum R_{X|Z}^2} = \frac{(\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z) R_{X|Z})^2}{\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z)^2 \sum R_{X|Z}^2}. \qquad (2.10)$$

Etape III : Répétition de l'algorithme et décision :

(1°) on répète l'étape II N fois (N grand) pour obtenir la loi de la statistique du test r_F^2 ;

(2°) on ordonne les valeurs de r_F^2 pour obtenir $(r_{F_{(1)}}^2, \cdots, r_{F_{(N)}}^2)$;

(3°) le test rejette H_0 si $r_{\text{obs}}^2 \geq r_{F_{(k)}}^2$ où $r_{F_{(k)}}^2$ est le quantile qui correspond à $k = [M(1-\alpha)]$, la partie entière de $M(1-\alpha)$.

2.3.3 L'approche de Kennedy

Cette technique de permutation a été proposée auparavant par des chercheurs dans un contexte multivarié pour tester le coefficient corrélation particlle pour des matrices de distances (Smouse et *al* 1986, Legendre et Fortin 1989).

La méthode est basée sur l'idée générale que le coefficient de corrélation partielle entre Y et X sachant Z, $r_{YX.Z}$, est égal au coefficient de corrélation simple entre les erreurs des deux modèles linéaires simples, $Y = \lambda_Y Z + \epsilon_{YZ}$ et $X = \lambda_X Z + \epsilon_{XZ}$ (voir démonstration, section 2.4). Ce test correspond au premier test exact présenté à la section 2.3.1.

Comme sous H_0 : $\beta_1 = 0$ (modèle réduit de 2.4) la statistique du test est le carré du coefficient de corrélation simple entre les résidus $R_{X|Z}$ et $R_{Y|Z}$, on permute ici les résidus de la régression de Y sur Z pour obtenir $R_{Y|Z}^{\pi}$ et on calcule la corrélation entre $R_{X|Z}$ et $R_{Y|Z}^{\pi}$ pour chaque permutation π.

C'est seulement le numérateur qui change à travers les permutations, car chaque permutation donne un nouvel appariement entre les deux résidus $(R_{Y|Z}^{\pi}, R_{X|Z})$ alors que les

sommes des carrés des résidus ne change pas ($\sum R^2_{Y|Z} = \sum (R^{\pi}_{Y|Z})^2$). Les étapes d'un test de permutation en utilisant la méthode de Kennedy (1995) sont :

Etape I · Calcul de la statistique observée :

identique à l'étape I de la méthode de Freedman et Lane.

Etape II : Algorithme de permutation :

(1)° on permute aléatoirement les résidus $R_{Y|Z}$ pour obtenir le vecteur $R^{\pi}_{Y|Z}$;

(2)° on recalcule la statistique du test, r^2_K, qui est le carré de la corrélation simple entre $R^{\pi}_{Y|Z}$ et $R_{X|Z}$;

$$r^2_K = \frac{(\sum R^{\pi}_{Y|Z} R_{X|Z})^2}{\sum R^2_{Y|Z} \sum R^2_{X|Z}}. \tag{2.11}$$

Etape III : Répétition de l'algorithme et décision :

on refait l' étape III décrite dans l'approche de Freedman et Lane, où r^2_K remplace r^2_F.

Notons que cette approche demande moins de calcul que celle de Freedman et Lane (1983) car la régression ne doit pas être refaite pour chaque permutation.

2.3.4 L'approche de Manly

Manly (1991) a proposé une approche différente pour faire le test de permutation. Pour obtenir la loi de la statistique du test il faut permuter la variable de réponse, Y. La justification de cette méthode repose sur la considération générale qui suggère que, si Y est indépendante de (X, Z), alors le mécanisme qui génère les données donne des probabilités égales à Y d'apparaître avec n'importe quel couple (X, Z). Cela signifie que la distribution des observations Y est similaire à celle des erreurs sous l'hypothèse nulle. On remarque ici que cette justification assume implicitement que $\beta_2 = 0$ (en plus de $\beta_1 = 0$), ce qui n'est pas toujours vrai.

Dans cette approche, on permute Y en gardant Z et X fixes et, pour chaque permutation, on calcule le coefficient de corrélation entre les résidus de la régression de X sur Z,

$R_{X|Z}$ et les résidus de la régression de Y^{π} sur Z, $R_{Y^{\pi}|Z}$. Les différentes étapes de cette procédure sont (l'indice M identifie Manly) :

Etape I : Calcul de la statistique observée :

identique à l'étape I de la méthode de Freedman et Lane.

Etape II : Algorithme de permutation :

(1)° les valeurs de Y sont permutées aléatoirement pour obtenir un autre vecteur Y^{π} ;

(2)° le vecteur Y^{π} est régressé sur Z pour obtenir les résidus $R_{Y^{\pi}|Z}$ avec $R_{Y^{\pi}|Z} = Y^{\pi} - \hat{\lambda}_{\pi(M)}Z$ où $\hat{\lambda}_{\pi(M)} = \sum Y^{\pi}Z / \sum Z^2$;

(3)° on recalcule la statistique du test, r^2_M, qui est le carré de la corrélation simple entre $R_{Y^{\pi}|Z}$ et $R_{X|Z}$;

$$r^2_M = \frac{(\sum R_{Y^{\pi}|Z} R_{X|Z})^2}{\sum R^2_{Y^{\pi}|Z} \sum R^2_{X|Z}}, \tag{2.12}$$

Etape III : Répétition de l'algorithme et décision :

on refait l' étape III décrite dans l'approche de Freedman et Lane, où r^2_M remplace r^2_F.

2.3.5 Comparaison entre les trois méthodes de permutation

Il faut d'abord noter qu'aucune des trois méthodes ne propose un test de permutation exact dans le sens défini au chapitre I. Anderson et Robinson (2001) ont montré que la méthode de Freedman et Lane (1983) donne le test le plus proche de leur test de permutation conceptuel exact (le deuxième test présenté à la section 2.3.1). Anderson et Legendre (1999) ont aussi montré que cette méthode donne les meilleurs résultats empiriques en termes d'erreur de type I et de puissance par rapport aux autres méthodes.

Même si Kennedy (1995) a prétendu que sa méthode donne des résultats identiques à celle de Freedman et Lane (1983), les deux méthodes sont differentes. En effet, la logique derrière les deux méthodes est semblable et se base sur le calcul des résidus sous le modèle

réduit (c'est à dire sous H_0). Puisque les vraies erreurs sont inconnues, on utilise les résidus pour les estimer. Cependant, les deux méthodes procèdent différemment pour générer la loi de la statistique de test sous permutation.

La différence entre les deux techniques de permutation est délicate, mais elle a des conséquences importantes. Pour la méthode de Kennedy, les résidus permutés $R^{\pi}_{Y|Z}$ sont utilisés directement dans le calcul de la statistique du test et ça veut dire que l'estimation du paramètre λ_Y reste fixe à travers les permutations.

Pour la méthode de Freedman et Lane (1983) les résidus $R^{\pi}_{Y|Z}$ sont utilisés pour créer de nouvelles observations $Y_{\pi(F)}$ qui seront régressées de nouveau sur la covariable Z afin de créer de nouveaux résidus $R_{Y_{\pi(F)}|Z}$. C'est à partir de ces résidus qu'on calcule la statistique du test.

Anderson et Robinson (2001) ont donné la relation entre les statistiques du test r^2_K, proposée par Kennedy (1995) et r^2_F, proposée par Freedman et Lane(1983). Cette relation est

$$r^2_F = \frac{r^2_K}{(1 - A^2_\pi)}, \tag{2.13}$$

où

$$A^2_\pi = \frac{(\sum R_{Y_\pi|Z}\, Z)^2}{\sum R^2_{Y_\pi|Z} \sum Z^2}$$

est le coefficient de corrélation entre $R_{Y_\pi|Z}$ et Z, avec $R_{Y_\pi|Z}$ les résidus de la régression de Y_π sur Z. La preuve détaillée de (2.13) est donnée à la section 2.4.

En effet, même si $\sum Z\, R_{Y|Z} = 0$, en permutant $R_{Y|Z}$ une petite relation entre Z et $R_{Y|Z}$ est réintroduite. Cela fait que A^2_π est différent de 0 pour toute permutation π. La relation (2.13) entre r^2_F et r^2_K montre que r^2_K est plus petite que r^2_F. Cela veut dire que la valeur observée apparaît plus extrême pour la méthode de Kennedy, ce qui donne une plus petite probabilité critique.

La méthode de Manly ne souffre pas du même problème que la méthode de Kennedy. Après chaque permutation des observations, Y^π est à nouveau régressée sur la covariable Z.

Par contre, même sous H_0, les triplets (x_i, z_i, y_i^π) ne sont pas équiprobables car l'espérance de Y_i dépend de $Z_i, i = 1, \cdots, n$. En effet, une étude empirique d'Anderson et Legendre (1999) a montré que l'approche de Manly est sensible aux données extrêmes de la covariable Z. Sous l'hypothèse $H_0 : \beta_1 = 0$, on a $Y = \beta_2 Z + \epsilon$. Alors, si les valeurs Z_i, $i = 1, \cdots, n$ ne sont pas trés différentes, on peut supposer que Y_i, $i = 1, \cdots, n$ sont approximativement de même loi. Dans ce cas, les unités permutées Y sont échangeables sous H_0. Mais, si on observe une valeur extrême de Z, disons Z^*, quand on permute les valeurs de la variable Y, la valeur Y^* de Y qui au départ correspondait à ce Z^* ne sera plus appariée avec Z^*. Finalement, on n'obtient pas un test précis de H_0.

2.3.6 Suggestion d'une nouvelle méthode

Dans nos simulations (chapitre III) nous avons exploré une approche différente mais proche de la méthode de Freedman et Lane (1983). L'idée était d'estimer le paramètre λ_Y, inconnu, à partir d'autres données, puis utiliser cette estimation et appliquer un test de permutation qui ressemble plus au test exact.

En principe, on pourrait faire ceci en mettant de coté une partie des observations, disons 20%, et l'utiliser pour estimer le paramètre inconnu. Après, on appliquerait le test conceptuel exact sur la partie restante. Par contre, si on procède ainsi, on réduit la taille de l'échantillon utilisé pour effectuer le test, et il serait difficile de comparer la performance d'une telle procédure avec les autres méthodes.

En considérant ceci, nous avons procédé d'une autre façon. On a effectué des tirages avec remise, de type « bootstrap» afin d'estimer le paramètre inconnu λ_Y et avons utilisé cette estimation pour faire le test «exact» de Anderson et Robinson (2001), à partir de toutes les données.

2.4 Quelques résultats théoriques

Dans cette section nous complétons quelques preuves présentées dans ce chapitre.

2.4.1 Relation entre la statistique de Freedman et Lane et la statistique de Kennedy

On va démontrer ici un énoncé de Anderson et Robinson (2001) sur la relation (2.13) entre la statistique du test proposé par Freedman et Lane (1983) et celle du test de Kennedy (1995). On a que

$$r_F^2 = \frac{r_K^2}{1 - A_\pi^2},$$

où

$$A_\pi^2 = \frac{(\sum R_{Y_\pi|Z}\, Z)^2}{\sum R_{Y_\pi|Z}^2 \sum Z^2}, \quad \text{et} \quad r_K^2 = \frac{(\sum R_{Y|Z}^\pi R_{X|Z})^2}{\sum R_{Y|Z}^2 \sum R_{X|Z}^2}.$$

Rappelons que la statistique de Freedman et Lane (1983) est :

$$r_F^2 = \frac{(\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z) R_{X|Z})^2}{\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z)^2 \sum R_{X|Z}^2}, \tag{2.14}$$

où

$$Y_{\pi(F)} = \hat{\lambda}_Y Z + R_{Y|Z}^\pi \quad \text{et} \quad \hat{\lambda}_{Y_{\pi(F)}} = \frac{\sum (Y_{\pi(F)} Z)}{\sum Z^2}.$$

On remplace $Y_{\pi(F)}$ dans la formule de $\hat{\lambda}_{\pi(F)}$ et on obtient :

$$\begin{aligned}
\hat{\lambda}_{\pi(F)} &= \frac{\sum (\hat{\lambda}_Y Z + R_{Y|Z}^\pi) Z}{\sum Z^2} \\
&= \frac{\sum (\hat{\lambda}_Y Z^2 + Z\, R_{Y|Z}^\pi)}{\sum Z^2} \\
&= \frac{\hat{\lambda}_Y \sum Z^2 + \sum Z\, R_{Y|Z}^\pi}{\sum Z^2} \\
&= \hat{\lambda}_Y + \frac{\sum Z\, R_{Y|Z}^\pi}{\sum Z^2}.
\end{aligned}$$

Ensuite, on réécrit la différence

$$\begin{aligned}
Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z &= \hat{\lambda}_Y Z + R_{Y|Z}^\pi - Z \left(\hat{\lambda}_Y + \frac{\sum Z\, R_{Y|Z}^\pi}{\sum Z^2} \right) \\
&= R_{Y|Z}^\pi - Z \frac{\sum Z\, R_{Y|Z}^\pi}{\sum Z^2}.
\end{aligned}$$

On remplace ce dernier résultat dans le numérateur et le dénominateur de la statistique de Freedman, r_F^2. Pour le numérateur on a :

$$\begin{aligned}
\left(\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z) R_{X|Z}\right)^2 &= \left(\sum \left(R^{\pi}_{Y|Z} - Z \frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2}\right) R_{X|Z}\right)^2 \\
&= \left(\sum (R^{\pi}_{Y|Z}\, R_{X|Z} - Z\, R_{X|Z} \frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2})\right)^2 \\
&= \left(\sum R^{\pi}_{Y|Z}\, R_{X|Z} - \underbrace{\sum Z\, R_{X|Z}}_{0} \frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2}\right)^2 \\
&= \left(\sum R^{\pi}_{Y|Z}\, R_{X|Z}\right)^2,
\end{aligned}$$

et on retrouve le numérateur de la statistique de Kennedy, r_K^2. Si on introduit $\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z)^2$ dans le dénominateur, on obtient :

$$\begin{aligned}
\sum \left(Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z\right)^2 &= \sum \left(R^{\pi}_{Y|Z} - Z \frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2}\right)^2 \\
&= \sum \left((R^{\pi}_{Y|Z})^2 - 2\, R^{\pi}_{Y|Z} Z \frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2} + \left(Z \frac{\sum Z R^{\pi}_{Y|Z}}{\sum Z^2}\right)^2\right) \\
&= \sum \left(R^{\pi}_{Y|Z}\right)^2 - 2 \sum R^{\pi}_{Y|Z} Z \frac{\sum Z R^{\pi}_{Y|Z}}{\sum Z^2} + \sum Z^2 \left(\frac{\sum Z\, R^{\pi}_{Y|Z}}{\sum Z^2}\right)^2 \\
&= \sum \left(R^{\pi}_{Y|Z}\right)^2 - 2 \frac{\left(\sum Z R^{\pi}_{Y|Z}\right)^2}{\sum Z^2} + \frac{\left(\sum Z R^{\pi}_{Y|Z}\right)^2}{\sum Z^2} \\
&= \sum \left(R^{\pi}_{Y|Z}\right)^2 - \frac{\left(\sum Z\, R^{\pi}_{Y|Z}\right)^2}{\sum Z^2}.
\end{aligned}$$

On remplace ces résultats dans la formule de r_F^2, et on obtient :

$$\begin{aligned} r_F^2 &= \frac{\left(\sum (Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z) R_{X|Z}\right)^2}{\sum \left(Y_{\pi(F)} - \hat{\lambda}_{\pi(F)} Z\right)^2 \sum R_{X|Z}^2} \\ &= \frac{\left(\sum R_{Y|Z}^{\pi} R_{X|Z}\right)^2}{\left(\sum \left(R_{Y|Z}^{\pi}\right)^2 - \frac{(\sum Z R_{Y|Z}^{\pi})^2}{\sum Z^2}\right) \sum R_{X|Z}^2} \\ &= \frac{\left(\sum R_{Y|Z}^{\pi} R_{X|Z}\right)^2}{\sum (R_{Y|Z}^{\pi})^2 \sum R_{X|Z}^2 - \sum R_{X|Z}^2 \frac{(\sum Z R_{Y|Z}^{\pi})^2}{\sum Z^2}}. \end{aligned}$$

En divisant le numérateur et le dénominateur avec $\sum R_{X|Z}^2 \sum (R_{Y|Z}^{\pi})^2$ on obtient le résultat, car

$$\begin{aligned} r_F^2 &= \frac{\frac{\left(\sum R_{Y|Z}^{\pi} R_{X|Z}\right)^2}{\sum R_{X|Z}^2 \sum (R_{Y|Z}^{\pi})^2}}{\frac{\sum (R_{Y|Z}^{\pi})^2 \sum R_{X|Z}^2 - \sum R_{X|Z}^2 \frac{(\sum Z R_{Y|Z}^{\pi})^2}{\sum Z^2}}{\sum R_{X|Z}^2 \sum (R_{Y|Z}^{\pi})^2}} \\ &= \frac{\frac{\left(\sum R_{Y|Z}^{\pi} R_{X|Z}\right)^2}{\sum R_{X|Z}^2 \sum (R_{Y|Z}^{\pi})^2}}{1 - \frac{(\sum Z R_{Y|Z}^{\pi})^2}{\sum Z^2 \sum (R_{Y|Z}^{\pi})^2}} \\ &= \frac{\frac{\left(\sum R_{Y|Z}^{\pi} R_{X|Z}\right)^2}{\sum R_{X|Z}^2 \sum R_{Y|Z}^2}}{1 - \frac{(\sum Z R_{Y|Z}^{\pi})^2}{\sum Z^2 \sum R_{Y|Z}^2}} \\ &= \frac{r_K^2}{1 - A_{\pi}^2}. \end{aligned}$$

2.4.2 Quelques résultats classiques utilisés dans ce chapitre

Dans les démonstrations suivantes on suppose, sans perte de généralité, qu'on a les trois variables X, Y et Z avec des espérances nulles. On commence par démontrer l'égalité (2.7).

Soit le vecteur aléatoire (X, Y, Z) ; alors, la corrélation partielle entre Y et X sachant Z se définit par

$$\rho_{XY.Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{YZ}^2}\sqrt{1-\rho_{XZ}^2}}. \tag{2.15}$$

Dans le cas où (X, Y, Z) suit une loi normale multivariée, $\rho_{XY.Z}$ est la corrélation entre Y et X conditionnellement à Z. En général, cette corrélation partielle se définit comme la corrélation entre ϵ_{YZ} et ϵ_{XZ} où

$$\epsilon_{XZ} = Y - \lambda_X Z, \quad \epsilon_{YZ} = Y - \lambda_Y Z,$$

et $\lambda_X Z$, respectivement $\lambda_Y Z$ est le meilleur approximant linéaire de X, respectivement Y (on suppose toujours que $\mu_X = \mu_Y = 0$). En d'autres mots, ces valeurs λ_Y et λ_X minimisent les quantités $E[(Y - \lambda_Y Z)^2$ et $E[(X - \lambda_X Z)^2]$. Après avoir calculé les dérivées par rapport à λ, on peut voir que

$$\lambda_X = \frac{E[XZ]}{E[Z^2]} = \frac{\mathrm{cov}(X,Z)}{\sigma_Z^2}, \qquad \lambda_Y = \frac{E[YZ]}{E[Z^2]} = \frac{\mathrm{cov}(Y,Z)}{\sigma_Z^2}.$$

De plus, on a $\mathrm{cov}(\epsilon_{XZ}, Z) = 0$ et $\mathrm{cov}(\epsilon_{YZ}, Z) = 0$. D'après la définition de la corrélation entre deux variables on a :

$$\rho_{\epsilon_{YZ}\epsilon_{XZ}} = \frac{\mathrm{cov}(\epsilon_{YZ}, \epsilon_{XZ})}{\sigma_{\epsilon_{YZ}}\sigma_{\epsilon_{XZ}}},$$

et on peut montrer que $\rho_{XY.Z} = \rho_{\epsilon_{YZ}\epsilon_{XZ}}$ où $\rho_{XY.Z}$ a été défini en (2.15) La covariance entre les erreurs est égale à

$$\begin{aligned}
\text{cov}(\epsilon_{YZ}, \epsilon_{XZ}) &= E[(X - \lambda_X Z)(Y - \lambda_Y Z)] \\
&= E[XY] - \lambda_X E[YZ] - \lambda_Y E[XZ] + \lambda_X \lambda_Y E[Z^2] \\
&= E[XY] - \frac{E[YZ]E[XZ]}{E[Z^2]} - \frac{E[YZ]E[XZ]}{E[Z^2]} + \frac{E[XZ]E[YZ]}{E[Z^2]} \\
&= E[XY] - \frac{E[YZ]E[XZ]}{E[Z^2]} \\
&= \text{cov}(X,Y) - \frac{\text{cov}(X,Z)\text{cov}(Y,Z)}{\sigma_Z^2},
\end{aligned}$$

tandis que la variance de l'erreur ϵ_{YZ} est égale à :

$$\begin{aligned}
\sigma^2_{\epsilon_{YZ}} &= E[(Y - \lambda_Y Z)^2] \\
&= E[(Y^2 - 2\lambda_Y YZ + \lambda_Y^2 Z^2)] \\
&= E[(Y^2)] - 2\frac{E[YZ]^2}{E[Z^2]} + \frac{E[YZ]^2}{E[Z^2]} \\
&= E[Y^2] - \frac{E[YZ]^2}{E[Z^2]} \\
&= \sigma_Y^2 - \frac{\text{cov}(Y,Z)}{\sigma_Z^2}.
\end{aligned}$$

De la même façon, on calcule la variance de ϵ_{XZ} et on trouve que $\sigma^2_{\epsilon_{XZ}} = \sigma_X^2 - \frac{\text{cov}(X,Z)}{\sigma_Z^2}$. On remplace les valeurs de $\text{cov}(\epsilon_{YZ}, \epsilon_{XZ})$, $\sigma_{\epsilon_{XZ}}$ et $\sigma_{\epsilon_{YZ}}$ dans la formule de la corrélation simple et on obtient :

$$\begin{aligned}
\rho_{\epsilon_{YZ}\epsilon_{XZ}} &= \frac{\text{cov}(X,Y) - \frac{\text{cov}(X,Z)\text{cov}(Y,Z)}{\sigma_Z^2}}{\sqrt{\sigma_Y^2 - \frac{\text{cov}^2(Y,Z)}{\sigma_Z^2}}\sqrt{\sigma_X^2 - \frac{\text{cov}^2(X,Z)}{\sigma_Z^2}}} \\
&= \frac{\frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} - \frac{\text{cov}(X,Z)}{\sigma_X\sigma_Y}\frac{\text{cov}(Y,Z)}{\sigma_X\sigma_Y}}{\sqrt{1 - \frac{\text{cov}^2(Y,Z)}{\sigma_Y^2\sigma_Z^2}}\sqrt{1 - \frac{\text{cov}^2(X,Z)}{\sigma_X^2\sigma_Z^2}}} \\
&= \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}} = \rho_{XY\,Z},
\end{aligned}$$

tel que désiré.

Finalement, on démontre l'équivalence entre les hypothèses $H_0 : \beta_1 = 0$ et $\widetilde{H}_0 : \rho_{XY.Z} = 0$. Soit le modèle de régression multiple $Y = \beta_1 X + \beta_2 Z + \epsilon$. Sous l'hypothèse nulle, le modèle linéaire multiple devient $Y = \beta_2 Z + \epsilon$. Soit les covariances $\text{cov}(Y, X)$ et $\text{cov}(Y, Z)$ sous H_0, qui se calculent comme suit :

$$\begin{aligned} \text{cov}(X, Y) &= E[XY] \\ &= E[X(\beta_2 Z + \epsilon)] \\ &= \beta_2 E[XZ] + \underbrace{E[X\epsilon]}_{0} \\ &= \beta_2 \text{cov}(X, Z), \end{aligned}$$

ainsi que

$$\begin{aligned} \text{cov}(Y, Z) &= E[YZ] \\ &= E[Z(\beta_2 Z + \epsilon)] \\ &= \beta_2 E[Z^2] + \underbrace{E[Z\epsilon]}_{0} \\ &= \beta_2 \sigma_z^2 , \end{aligned}$$

car Z et ϵ ainsi que X et ϵ sont de covariance nulle par hypothèse. Cela donne $\beta_2 = \text{cov}(Y, Z)/\sigma_z^2$, ainsi que :

$$\begin{aligned} \rho_{XY} &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{\beta_2 \text{cov}(X, Z)}{\sigma_X \sigma_Y} \\ &= \frac{\text{cov}(Y, Z)\text{cov}(X, Z)}{\sigma_z^2} \frac{1}{\sigma_X \sigma_Y} \\ &= \frac{\text{cov}(Y, Z)}{\sigma_Y \sigma_z} \frac{\text{cov}(X, Z)}{\sigma_X \sigma_z} \\ &= \rho_{YZ} \rho_{XZ} . \end{aligned}$$

Comme le numérateur de la corrélation partielle est égal à $\rho_{XY} - \rho_{YZ}\rho_{XZ}$, on conclut d'après le résultat précédent que β est nul si est seulement si la corrélation partielle est nulle.

CHAPITRE III

COMPARAISON EMPIRIQUE DE QUATRE MÉTHODES DE PERMUTATION

Dans ce chapitre, on compare, en utilisant des simulations, l'erreur de type I et la puissance empiriques des tests de permutation approximatifs présentés dans le chapitre II . On examine l'effet des paramètres suivants :

(i) la taille de l'échantillon n ;

(ii) le degré de colinéarité entre les covariables, ρ_{XZ} ;

(iii) la valeur du coefficient de la covariable Z, β_2 ;

(iv) la loi de l'erreur aléatoire, ϵ.

3.1 Les méthodes de simulation

Les simulations ont été faites à travers plusieurs niveaux de tous les facteurs à l'étude (voir section 3.1.1). Le programme informatique utilisé pour les simulations et les calculs a été écrit et exécuté dans le langage R. L'étude a été faite à partir de 2000 séries de données simulées pour chaque combinaison de facteurs et 999 permutations pour chaque test effectué.

3.1.1 Les facteurs à l'étude

Notre modèle de régression multiple est $Y_i = \beta_1 X_i + \beta_2 Z_i + \epsilon_i$, $i = 1, \cdots, n$, et l'hypothèse nulle d'intérêt est $H_0 : \beta_1 = 0$. C'est à dire, sous H_0 il n'y a aucun effet significatif de

la variable X dans la régression linéaire multiple. Les données ont été simulées pour toutes les combinaisons des niveaux choisis pour chacun des facteurs. On a simulé des échantillons en tenant compte des différents niveaux des facteurs suivants :

- Puisqu'on veut connaître le comportement des quatre méthodes selon la taille de l'échantillon on a considéré différentes tailles (n «petit» et n «grand») : $n \in \{10, 30, 45, 60\}$.
- On a considéré que les variables explicatives sont fixes. Donc, pour chaque taille n on a généré une seule fois un couple de loi bivariée normale (X_i, Z_i), $i = 1, \cdots, n$.
- Les quatre méthodes permettent la colinéarité (c-à-d la corrélation linéaire entre les variables explicatives ρ_{XZ} peut être différente de 0) et c'est pour cela qu'on a considéré deux situations, indépendance et de forte relation linéaire et on a pris : $\rho_{XZ} \in \{0, 0.9\}$.
- Puisque les tests de permutation nécessitent l'échangeabilité seulement sous H_0, on a pris des situations où les suppositions du modèle linéaire ne sont pas verifiées (la normalité et l'indépendance des erreurs). Pour la violation de la condition de normalité, on a choisi des erreurs qui ne sont pas normales, mais qui suivent une loi centrée sur 0, $\epsilon_i \sim Uniforme(-1, 1)$. Pour la violation de l'indépendance on a pris des erreurs de loi normale dépendantes et telles que la corrélation $\rho_{\epsilon_i \epsilon_j}$, $i \neq j$ est différente de 0. On a considéré un vecteur $(\epsilon_1, \cdots, \epsilon_n)$ qui suit une loi normale multivariée $\underline{\epsilon} = (\epsilon_1, \cdots, \epsilon_n) \sim N(0, \Sigma)$ (afin d'avoir l'échangebilité).

  Pour la valeur du coefficient de la covariable Z on a pris deux valeurs du coefficient de régression : $\beta_2 \in \{0, 1.5\}$.

Notons que, pour introduire la colinéarité entre les covariables X et Z, on génère un vecteur de loi normale bivariée $(X, Z) \sim N(0, \Sigma)$ tel que X et Z sont dépendantes avec une corrélation $\rho_{XZ} \neq 0$. Pour cela on procède par les étapes suivantes :

1) On génère deux variables X_1 et Z_1 de loi normale centrée réduite $X_1, Z_1 \sim N(0, 1)$ indépendantes.

2) On calcule $\Sigma^{\frac{1}{2}}$ à partir de la matrice de variance-covariance où :

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}.$$

Pour créer $\Sigma^{\frac{1}{2}}$, on fait la décomposition orthogonale de la matrice de variance-covariance. On calcule les valeurs propres (λ_1, λ_2) de Σ et la matrice orthogonale P constituée de ses vecteurs propres V_1, V_2. Alors, la matrice $\Sigma^{\frac{1}{2}}$ est égale à

$$\Sigma^{\frac{1}{2}} = P \Lambda^{\frac{1}{2}} P^T,$$

où

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix}.$$

3) Finalement, on calcule les nouvelles valeurs des variables X et Z corrélées en utilisant l'équation suivante

$$(X, Z) = \Sigma^{\frac{1}{2}}(X_1, Z_1). \tag{3.1}$$

Alors, on peut voir que la matrice de variance covariance de (X, Z) est en effet égale à Σ. Notons que, contrairement à l'étude de Anderson et Legendre (1999), on exclut le choix des erreurs de loi exponentielle cubique $[Exp(1)]^3$ car elle n'est pas centrée sur 0 et cela contredit la notion d'erreur.

3.1.2 Étude de l'erreur type I

Pour l'étude de l'erreur de type I, on a fait les simulations lorsque l'hypothèse nulle est vraie $H_0 : \beta_1 = 0$ en considérant les différentes combinaisons des facteurs cités dans la section précédante. On a fait 999 permutations pour chaque ensemble de données simulées et on a considéré un seuil de signification égal à $\alpha = 0.05$. L'erreur de type I empirique a été calculée pour les quatre tests de permutation et le test t de Student. Cette erreur de type I empirique est la proportion des probabilités critiques inferieures à α, donc se calcule par la formule suivante ;

$$\text{Erreur type I empirique} = \frac{(N^0 \text{ p-valeur } \leq \alpha)}{2000}. \tag{3.2}$$

3.1.3 Étude de la puissance

Dans ce cas, la puissance a été définie par la proportion de rejets parmi les 2000 simulations lorsque l'hypothèse nulle est fausse, c-à-d sous $H_1 : \beta_1 \neq 0$. Pour la calculer, on a utilisé la formule suivante,

$$\text{Puissance empirique} = \frac{(N^0 \text{ p-valeur } \leq \alpha)}{2000}. \tag{3.3}$$

Pour l'étude de la puissance, les données ont été générées de la même manière, sauf qu'on a pris des valeurs de $\beta_1 \neq 0$. Le paramètre β_1 a été choisi tel que la corrélation partielle soit égale à une valeur donnée, et on a pris : $\rho_{XY.Z} = 0.5$ (valeur «moyenne » pas trop grande). À partir des formules appropriées on peut déduire la valeur de β_1 en fonction des autres paramètres, $\beta_1 = f(\rho_{YX\,Z}, \rho_{XZ}, \beta_2)$.

En effet, voila comment on peut obtenir β_1. Soit X et Z deux variables normales centrées réduites ($\sigma_X = \sigma_Z = 1$) avec une corrélation $\rho_{XZ} = \text{cov}(X,Z)/\sigma_X^2\sigma_Y^2 = \text{cov}(X,Z)$. D'après le modèle de régression linéaire multiple on a $Y = \beta_1 X + \beta_2 Z + \epsilon$, d'où la variance de Y, $V(Y) = \sigma_Y^2$ est égale à :

$$\begin{aligned} V(Y) &= V(\beta_1 X + \beta_2 Z + \epsilon) \\ &= \beta_1^2 V(X) + \beta_2^2 V(Z) + \beta_1\beta_2 \text{cov}(X,Z) + V(\epsilon) \\ &= \beta_1^2 + \beta_2^2 + \beta_1\beta_2\rho_{XZ} + \sigma_\epsilon^2. \end{aligned}$$

Les autres termes de covariance sont nuls, car ϵ est supposé tel que $\text{cov}(X,\epsilon) = 0$ et $\text{cov}(Z,\epsilon) = 0$. On calcule maintenant les expressions des différentes covariances pour les remplacer dans la formule de la corrélation partielle. Soit

$$\begin{aligned} \text{cov}(Y,Z) &= \text{cov}(\beta_1 X + \beta_2 Z + \epsilon, Z) \\ &= \beta_1 \underbrace{\text{cov}(X,Z)}_{\rho_{XZ}} + \beta_2 \underbrace{V(Z)}_{1} + \underbrace{\text{cov}(Z\epsilon)}_{0} \\ &= \beta_1\rho_{XZ} + \beta_2, \end{aligned}$$

ainsi que

$$\begin{aligned}\operatorname{cov}(Y,X) &= \operatorname{cov}(\beta_1 X + \beta_2 Z + \epsilon, X) \\ &= \beta_1 \underbrace{V(X)}_{1} + \beta_2 \underbrace{\operatorname{cov}(X,Z)}_{\rho_{XZ}} + \underbrace{\operatorname{cov}(X\epsilon)}_{0} \\ &= \beta_1 + \beta_2 \rho_{XZ}.\end{aligned}$$

On remplace ces résultats dans $\rho_{XY.Z}$, en tenant compte du fait que $\sigma_X = \sigma_Z = 1$:

$$\begin{aligned}\rho_{XY.Z} &= \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)}\sqrt{(1-\rho_{YZ}^2)}} \\ &= \frac{\dfrac{\operatorname{cov}(X,Y)}{\sigma_Y\sigma_X} - \dfrac{\operatorname{cov}(X,Z)}{\sigma_X\sigma_Z}\dfrac{\operatorname{cov}(Y,Z)}{\sigma_Y\sigma_Z}}{\sqrt{1-\dfrac{\operatorname{cov}^2(X,Z)}{\sigma_X^2\sigma_Z^2}}\sqrt{1-\dfrac{\operatorname{cov}^2(Y,Z)}{\sigma_Y^2\sigma_Z^2}}} \\ &= \frac{\operatorname{cov}(X,Y) - \operatorname{cov}(X,Z)\operatorname{cov}(Y,Z)}{\sqrt{1-\operatorname{cov}(X,Z)^2}\sqrt{\sigma_Y^2 - \operatorname{cov}(Y,Z)^2}} \\ &= \frac{\beta_1 + \beta_2\rho_{XZ} - \rho_{XZ}(\beta_1\rho_{XZ} + \beta_2)}{\sqrt{1-\rho_{XZ}^2}\sqrt{(\beta_1^2 + \beta_2^2 + \beta_1\beta_2\rho_{XZ} + \sigma_\epsilon^2 - (\beta_1\rho_{XZ} + \beta_2)^2)}} \\ &= \frac{\beta_1(1-\rho_{XZ}^2)}{\sqrt{1-\rho_{XZ}^2}\sqrt{\sigma_\epsilon^2 - \beta_1\beta_2\rho_{XZ} + \beta_1^2(1-\rho_{XZ}^2)}} \\ &= \frac{\beta_1}{\sqrt{\beta_1^2 + \dfrac{\sigma_\epsilon^2 - \beta_1\beta_2\rho_{XZ}}{1-\rho_{XZ}^2}}}.\end{aligned}$$

D'après le calcul précédent on peut déduire les valeurs de β_1 correspondant à une corrélation partielle $\rho_{XY.Z} = 0.5$, si β_2, ρ_{XZ} et σ_ϵ^2 sont connus. Cela revient à résoudre l'équation de $2^{ème}$ degré suivante en β_1 :

$$\beta_1^2(1-\rho_{XY.Z}^2) + \beta_1\frac{\beta_2\rho_{XZ}\rho_{XY.Z}^2}{1-\rho_{XZ}^2} - \frac{\rho_{XY.Z}^2\sigma_\epsilon^2}{1-\rho_{XZ}^2} = 0. \tag{3.4}$$

Pour des erreurs de loi normale on a $\sigma_\epsilon^2 = 1$, et pour des erreurs de loi uniforme $[-1,1]$ on a $\sigma_\epsilon^2 = (b-a)^2/12 = 1/3$. Le tableau suivant présente les solutions de l'équation (3.4) selon les valeurs de ρ_{XYZ}, ρ_{XZ}, β_2 et σ_ϵ^2.

Tableau 3.1 Les solutions de l'équation (3.4) en fonction de la variance de l'erreur aléatoire σ_ϵ^2, de β_2 et ρ_{XZ}.

Loi de ϵ	ρ_{xz}	β_2	solution1	solution2
Normale	0	$\forall\beta_2$	0.58	-0.58
Uniforme	0	$\forall\beta_2$	0.19	-0.19
Normale	0.9	0	1.32	-1.32
Uniforme	0.9	0	0.44	-0.44
Normale	0.9	1.5	0.59	-2.95
Uniforme	0.9	1.5	2.45	-0.08

Puisqu'on a choisi la corrélation partielle $\rho_{XY.Z}$=0.5 positive, et on sait que $\rho_{XY.Z}$ et β_1 sont de même signe, on a retenu les solutions positives de l'équation (3.4) pour faire les simulations.

3.2 Résultats des simulations

Dans cette section, on présente les résultats des simulations. Les figures 3.1 et 3.2 comparent les quatre méthodes approximatives de test de permutation présentés dans le chapitre précédant, ainsi que le test t de Student en terme d'erreur de type I empirique. En abscisse, on a la taille de l'échantillon, et chaque graphique correspond à une autre valeur de la colinéarité ρ_{XZ} et du coefficient de la covariable β_2. Les figures 3.3 et 3.4 comparent la puissance empirique des quatre méthodes et le test t de Student en fonction de la taille de l'échantillon, selon la colinéarité ρ_{XZ} et la valeur du coefficient de la covariable β_2. La méthode qu'on a proposé a donné des résultats identiques à la méthode de Freedman et Lane (1983) pour la majorité des simulations avec une seule différence minime, et c'est pourquoi on ne l'a pas présentée dans les graphiques.

3.2.1 Erreur de type I empirique

D'aprés les figures 3.1 et 3.2, on remarque que la méthode de Manly (1997) et Freedman et Lane (1983) donnent des résultats similaires pour les différentes situations.

Les résultats des simulations ont montré que la méthode de Kennedy donne une erreur type I empirique plus élevée que les autres méthodes et largement supérieure à $\alpha = 0.05$, lorsque la taille de l'échantillon est petite ($n = 10$). Ce problème diminue au fur et à mesure que la taille de l'échantillon augmente (figures 3.1 et 3.2). L'étude de Anderson et Legendre (1999) a montré que ce problème d'erreur de type I gonflée devient plus grave avec l'augmentation du nombre de covariables dans la régression multiple.

La présence du paramètre $\beta_2 \neq 0$ de la covariable Z n'a pas un effet remarquable sur l'erreur de type I pour les quatre méthodes de permutation (figure 3.1 c,d et figure 3.2 c,d). La présence de la colinéarité (dépendance linéaire) entre les covariables X. et Z, $\rho_{XZ} \neq 0$, a peu d'effet sur les résultats (figure 3.1 b,d et figure 3.2 b,d).

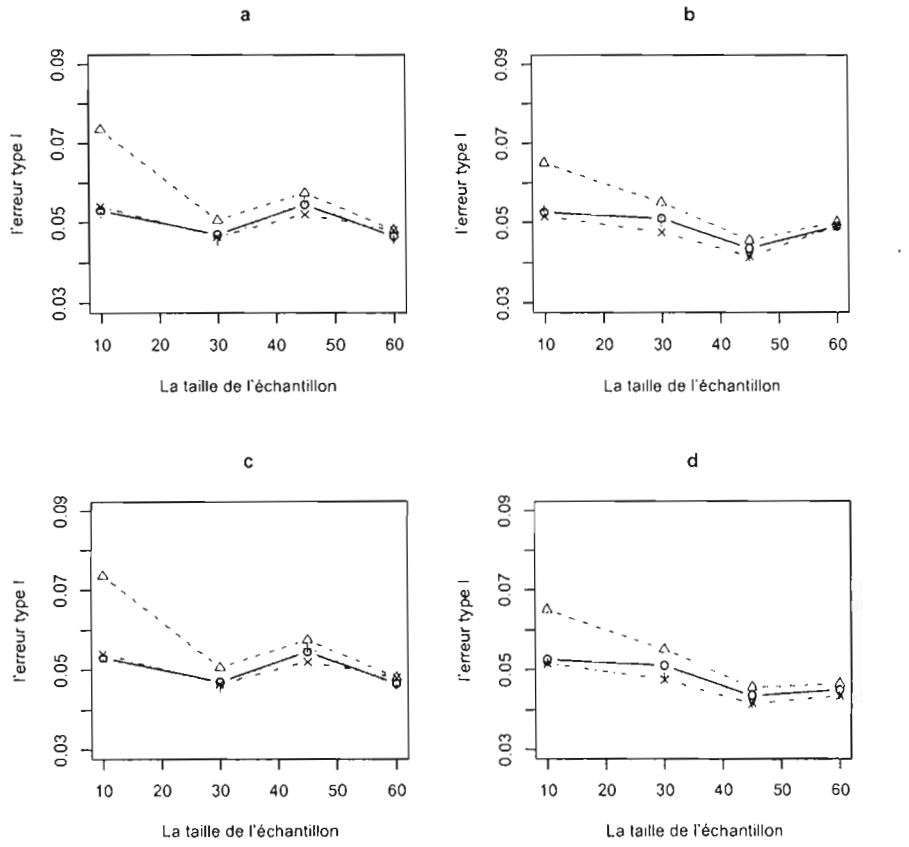
Lorsque les erreurs sont normales dépendantes, la différence entre la méthode de Kennedy (1995) et les autres méthodes est moins importante, même si l'erreur de type I reste encore gonflée. Les autres méthodes, Manly (1997), Freedman et Lane (1983) et la méthode proposée, donnent une erreur de type I qui se rapproche du résultat du test t de Student et ne diffèrent pas beaucoup de $\alpha = 0.05$. L'étude de Anderson et Legendre (1999) a trouvé que ces différentes méthodes de permutation convergent plus vite que le test t de Student vers l'erreur de type I appropriée si on considère des situations avec des erreurs extrêmement non normales $[exp(1)]^3$ (choix exclu dans nos simulations).

Contrairement aux résultats de l'étude de Anderson et Legendre (1999), le test t de Student donne des erreurs de type I très bonnes, mais pas toujours inférieures à $\alpha = 0.05$.

À l'exception de la méthode de Kennedy, les autres méthodes de permutation donnent des résultats plus proches du test t de Student (surtout la méthode de Manly, 1997) pour les différentes combinaisons. Cela peut être expliqué par le fait qu'on a gardé les

variables X et Z fixes à travers les simulations et la méthode de Manly est supposée bien fonctionner dans cette situation. Les résultats de l'étude de Anderson et Legendre (1999) ont montré que la méthode de Manly donne une erreur de type I biaisée en présence de valeurs extrêmes dans les données, ce qui n'est pas notre cas.

Figure 3.1 Comparaison de l'erreur de type I de quatre tests de permutation et du test t de Student comme fonction de la taille d'échantillon, selon quatre valeurs de (β_2, ρ_{XZ}), lorsque les erreurs aléatoires suivent une loi uniforme $[-1, 1]$.

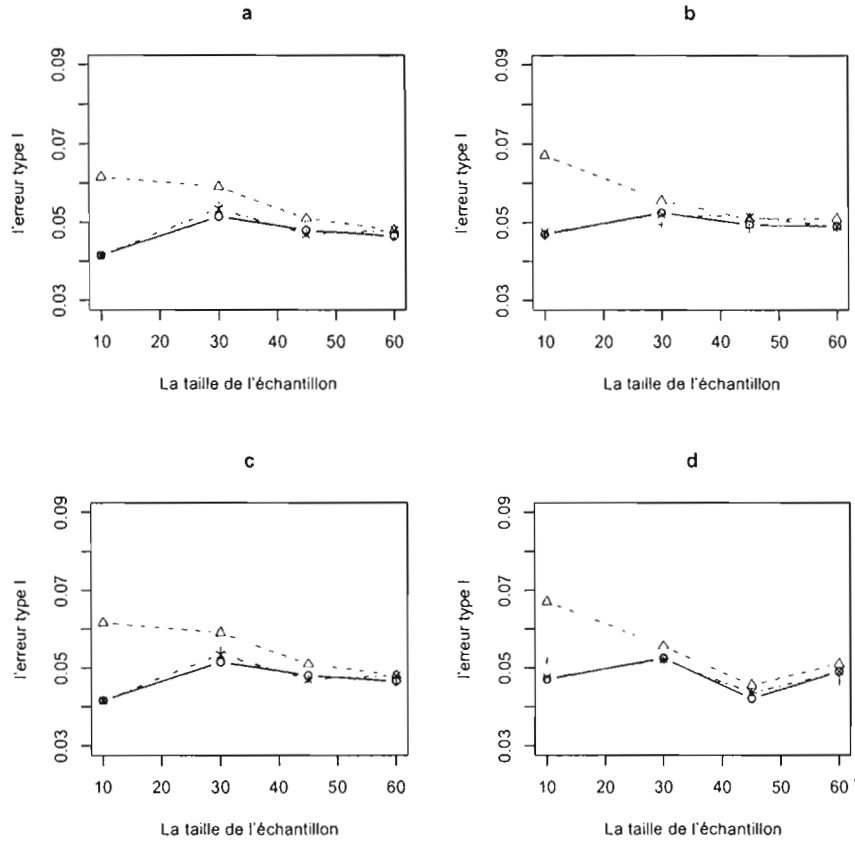


∘ Freedman et Lane ; △ Kennedy ; + Manly ; × test t de Student

a : $\beta_2 = 0$, $\rho_{XZ} = 0$; b : $\beta_2 = 0$, $\rho_{XZ} = 0.9$

c : $\beta_2 = 1.5$, $\rho_{XZ} = 0$; d : $\beta_2 = 1.5$, $\rho_{XZ} = 0.9$

Figure 3.2 Comparaison de l'erreur de type I de quatre tests de permutation et du test t de Student comme fonction de la taille d'échantillon, selon quatre valeurs de (β_2, ρ_{XZ}), lorsque les erreurs aléatoires sont de lois normales dépendantes.



∘ Freedman et Lane ; △ Kennedy ; + Manly ; × test t de Student

a : $\beta_2 = 0$, $\rho_{XZ} = 0$; b : $\beta_2 = 0$, $\rho_{XZ} = 0.9$

c : $\beta_2 = 1.5$, $\rho_{XZ} = 0$; d : $\beta_2 = 1.5$, $\rho_{XZ} = 0.9$

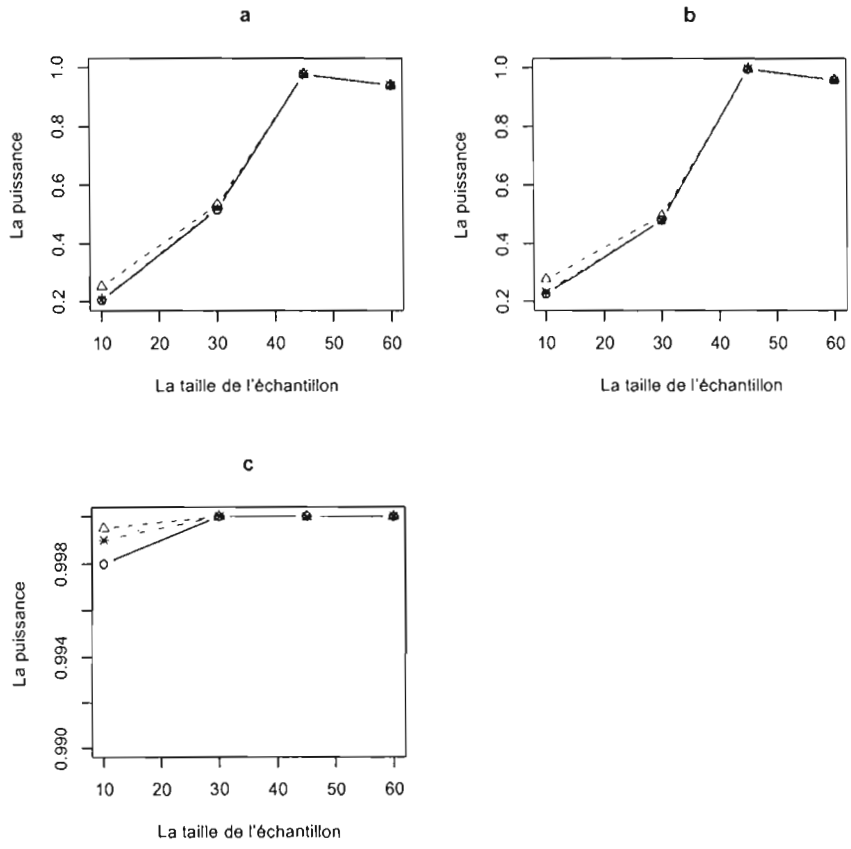
3.2.2 La puissance empirique

Pour l'étude de la puissance on a considéré des valeurs de β_2 qui correspondent à une corrélation partielle de $\rho_{YX.Z} = 0.5$. On note ici que, contrairement à l'étude de Anderson et Legendre (1999), on a inclus la méthode de Kennedy dans l'étude de la puissance. Ces résultats sont présentés dans les figures fig 3.3 et fig 3.4.

Toutes les méthodes montrent une augmentation de la puissance avec l'augmentation de la taille d'échantillon et l'augmentation de la valeur de β_1. Cela est naturel puisqu'on s'attend a ce que la puissance augmente avec l'augmentation de la taille d'échantillon. Aussi, tel qu'attendu, plus β_1 s'éloigne de 0, plus les tests sont puissants et capables de detecter que l'hypothèse nulle est fausse. On remarque que la méthode de Kennedy donne une puissance plus élevée que les autres méthodes, et que cette différence est plus importante lorsque la taille d'échantillon est petite, mais disparaît avec l'augmentation de la taille d'échantillon.

Pour ce qui est de la loi du terme aléatoire, on remarque que, pour les quatre tests de permutation et le test t de Student, la puissance est moins élevée lorsque les erreurs sont uniformes, que lorsqu'elles sont normales dépendantes, mais cette différence disparaît avec l'augmentation de la taille d'échantillon.

Figure 3.3 Comparaison de la puissance de quatre tests de permutation et du test t de Student comme fonction de la taille d'échantillon, selon quatre valeurs de (β_2, ρ_{XZ}), lorsque les erreurs aléatoires suivent une loi uniforme $[-1, 1]$.



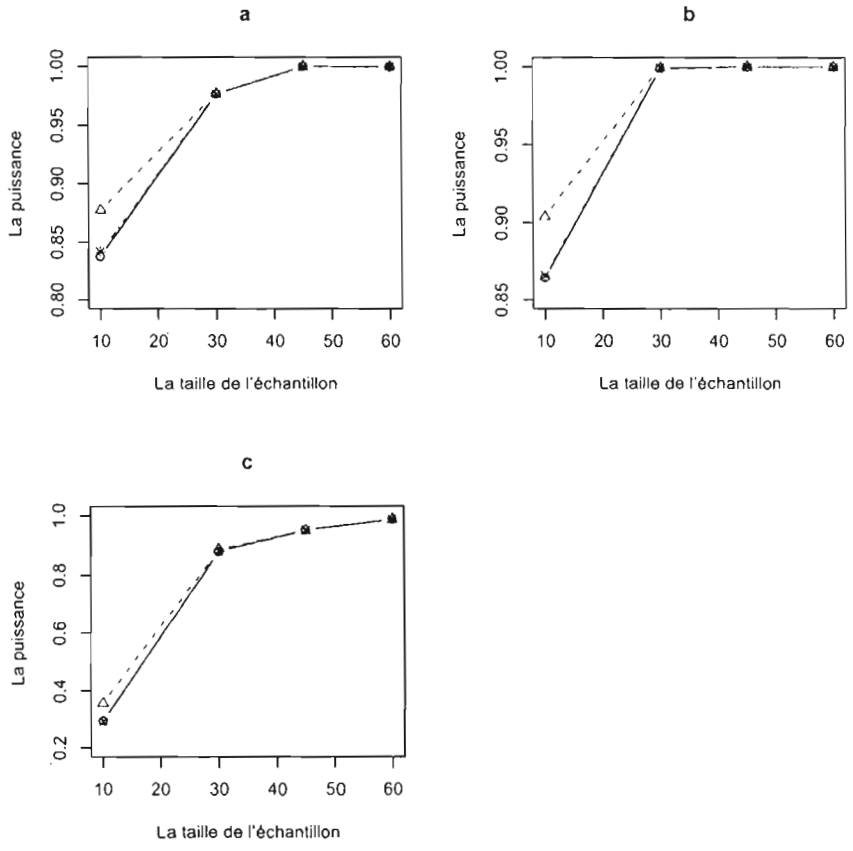
∘ Freedman et Lane ; △ Kennedy ; + Manly ; × test t de Student

a : Loi uniforme, $\beta_1 = 0.19$, $\beta_2 = 0$, $\rho_{XZ} = 0$;

b : Loi uniforme, $\beta_1 = 0.44$, $\beta_2 = 0$, $\rho_{XZ} = 0.9$;

c : Loi uniforme, $\beta_1 = 2.45$, $\beta_2 = 1.5$, $\rho_{XZ} = 0.9$.

Figure 3.4 Comparaison de la puissance de quatre tests de permutation et du test t de Student comme fonction de la taille d'échantillon, selon quatre valeurs de (β_2, ρ_{XZ}), lorsque les erreurs aléatoires sont de lois normales dépendantes.



∘ Freedman et Lane ; △ Kennedy ; + Manly ; × test t de Student

a : Normales dépendantes $\beta_1 = 0.57$, $\beta_2 = 1.5$, $\rho_{XZ} = 0$;

b : Normales dépendantes $\beta_1 = 1.32$, $\beta_2 = 0$, $\rho_{XZ} = 0.9$;

c : Normales dépendantes $\beta_1 = 0.59$, $\beta_2 = 1.5$, $\rho_{XZ} = 0.9$.

3.3 Conclusion

Les résultats obtenus par ces simulations montrent que la méthode de permutation proposée par Kennedy (1995) ne donne pas des résultats équivalents à la méthode proposée par Freedman et Lane (1983), en utilisant la corrélation partielle comme statistique de test. La méthode de Kennedy (1995) donne une erreur de type I gonflée, particulièrement lorsque la taille d'échantillon est petite ($n = 10$). L'étude de Anderson et Legendre (1999) a montré que ce problème d'erreur de type I devient plus sérieux avec l'augmentation du nombre de variables explicatives incluses dans le modèle. Donc la méthode de Kennedy (1995) ne donne pas les mêmes résultats que la méthode de Freedman et Lane (1983). Notons que Kennedy (1995) avait proposé sa méthode comme une alternative équivalente à la méthode de Freedman et Lane (1983) qui serait plus simple quant au niveau du calcul.

Anderson et Robinson (2001) ont montré théoriquement quelle est la différence entre ces deux méthodes de permutation (Kennedy, 1995 ; et Freedman et Lane, 1983). Cette formule a été présentée dans la section (2.4) du chapitre II. Nos simulations confirment cette différence entre ces deux méthodes.

Si on applique la même technique de permutation pour les tests de corrélation partielle entre les matrices de distances proposée par Smoose et al (1986) comme extension du test de Mantel dans le cas de plusieurs matrices qu'on va présenter dans le chapitre IV, on s'attend à ce qu'elle souffre des mêmes faiblesses que la méthode de Kennedy.

Pour ce qui est des trois autres méthodes, la permutation des valeurs de la variable dépendante proposée par Manly (1997) et la méthode de permutation des résidus sous le modèle réduit proposée par Freedman et Lane (1983), ainsi que la méthode qu'on a proposé (où le bootstrap est utilisé pour estimer le coefficient de régression inconnu) donnent des résultats équivalents dans la plupart des situations simulées.

Pour ce qui est de l'effet de la loi du terme aléatoire, et contrairement à l'étude de Anderson et Legendre (1999), on n'a pas détecté des différences importantes dans la

performances des quatres méthodes. Cela est dû peut être au fait qu'on n'a pas choisi des situations où il y a des lois très asymétriques et non centrées à 0. Les lois utilisées pour la génération des erreurs aléatoires sont symétriques ou se rapprochent de la loi normale.

Pour des tailles d'échantillon suffisamment grandes ($n \geq 45$) les quatre méthodes sont équivalentes pour ce qui est de l'erreur de type I et de la puissance. Dans un tel cas, on choisit celle qui demande moins de calcul. Dans un tel cas, on peut recommander d'utiliser la méthode de Kennedy (1995) puisqu'elle ne demande pas de refaire la régression après chaque permutation.

CHAPITRE IV

LE TEST DE MANTEL SIMPLE ET LE TEST DE MANTEL PARTIEL

Dans ce chapitre, on va présenter une généralisation des tests de permutation pour les coefficients partiels de régression à des matrices de distances. Certains tableaux de données multivariées portent sur les ressemblances (des similarités ou des dissimilarités). En effet, plusieurs méthodes d'analyse multivariée comme la classification (clustering) et l'ordination sont basées sur de telles ressemblances. Ces méthodes sont généralement utilisées dans un contexte exploratoire et descriptif des données, sans tests statistiques formels sur les relations entre les variables (Shannon, 2002).

Une dissimilarité (similarité) entre deux objets mesure à quel point ils sont différents (semblables). Parfois, on peut employer une distance entre les objets, mais une dissimilarité n'est pas nécessairement une distance. En mathématique, on appelle distance sur un ensemble E une application $D : E \times E \longrightarrow R$ telle que, pour tout $i, j, k \in E$:

- $D(i,j) \geq 0$, la distance est positive entre deux points ;
- $D(i,i) = 0$, la distance est zéro d'un point à lui même ;
- $D(i,j) = D(j,i)$, la distance est symétrique ;
- $D(i,k) \leq D(i,j) + D(k,j)$, inégalité triangulaire.

Souvent, seulement les trois premiers axiomes d'une distance sont satisfaits. Dans ce qui suit, sans perte de généralité, on considère seulement des matrices de distances symétriques.

Les chercheurs s'intéressent parfois à comparer deux ou plusieurs matrices de distances (entre plusieurs groupes de variables) sur les mêmes sujets, pour examiner l'hypothèse d'une présumée relation entre ces matrices de distances.

Plusieurs auteurs ont proposé des calculs simples d'un test de signification pour le coefficient de corrélation entre les éléments de deux matrices de distances. Cependant, les méthodes paramétriques comme la régression et la corrélation ne peuvent être appliquées. Les suppositions de la normalité et de l'indépendance nécessaires pour faire le test ne sont pas valides dans le cas des mesures de distance (Dietz, 1983). Dans la littérature, on propose des tests de permutation afin de surmonter cette difficulté. Nous trouvons ce point discutable et nous reviendrons là-dessus plus tard.

Une alternative simple et flexible a été proposée par Mantel (1967). Manly a nommé cette procédure le «test de Mantel», et elle est largement utilisée dans différents domaines de la recherche comme l'environnement, la biologie, l'écologie et la génétique pour comparer deux matrices de distances.

Dans ce chapitre, on va d'abord présenter de façon informelle la procédure introduite par Mantel, ainsi que les modifications proposées par Smoose et al. (1986). Ensuite, on décrit le test de Mantel partiel lorsqu'il y a plus de deux matrices de distances. Finalement, on présente un exemple d'application du test de Mantel simple et du test de Mantel partiel sur des données présentées par Manly (1997).

4.1 Le test de Mantel simple

4.1.1 Présentation du test

En 1967, Mantel a présenté une nouvelle approche pour évaluer la relation entre deux matrices de correspondances mesurées sur les mêmes individus. Plus tard, Manly (1986) a nommé cette approche «le test de Mantel». Pour bien comprendre ce test on va le présenter comme Mantel l'a fait dans son article de 1967. Ensuite, on va expliquer les modifications apportées à ce test par Smoose et al. (1986).

Mantel a proposé cette méthode dans le contexte de la détection d'un regroupement (clustering) en temps et en espace pour la maladie de leucémie pour $n(n-1)/2$ paires de distances formées à partir de n cas observés de cette maladie.

Supposons qu'on a n individus et que deux groupes de variables ont été mesurés pour chaque individu (dans notre exemple il s'agit de la variable temps et de la variable espace). Chaque groupe de variables est utilisé pour construire une matrice de distances. Ainsi, on a deux matrices de distances $\mathbf{X}_{n\times n}$ et $\mathbf{Y}_{n\times n}$. Par exemple, l'entrée de la $i^{ème}$ ligne et la $j^{ème}$ colonne de la matrice $\mathbf{X}$ mesure la distance entre les individus i et j basée sur le 1^{er} groupe de variables, tandis que l'entrée de la $i^{ème}$ ligne et la $j^{ème}$ colonne de la matrice $\mathbf{Y}$ mesure la distance entre les individus i et j basée sur le $2^{ème}$ groupe de variables.

Ces matrices de distances sont, dans la plupart des cas, symétriques et ont des éléments nuls sur la diagonale. En effet, la distance entre un individu et lui même est nulle, $X_{ii}=0$, et la distance entre l'individu i et l'individu j ne change pas, c'est-à-dire pour tout $i,j=1,\cdots,n$ $i\neq j$ $X_{ij}=X_{ji}$. Notons que la symétrie dépend de la métrique choisie pour calculer les distances. Dans ce cas, on a $N=\binom{n}{2}=n(n-1)/2$ couples distincts de distances (X_{ij},Y_{ij}), $i,j=1,\cdots,n$ ainsi que les matrices :

$$\mathbf{X}=\begin{pmatrix} 0 & X_{12} & \cdots & X_{1n} \\ X_{21} & 0 & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & 0 \end{pmatrix}, \mathbf{Y}=\begin{pmatrix} 0 & Y_{12} & \cdots & Y_{1n} \\ Y_{21} & 0 & \cdots & Y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & 0 \end{pmatrix}.$$

Mantel a proposé de faire une régression de l'une des matrices (temps) sur l'autre (espace) en mettant $\mathbf{Y}=\beta\mathbf{X}+\mathbf{E}$ où $\mathbf{E}$ est une matrice d'erreurs aléatoires, dans le but de détecter des regroupements (clusters), et non pas de vérifier une éventuelle relation entre les deux matrices.

En langage courant, l'hypothèse nulle du test est H_0 : il n'y a pas de regroupement «no clustering». En termes de $\mathbf{X}$,$\mathbf{Y}$ cela veut dire que les éléments correspondants des deux matrices ne sont pas associés. En d'autres mots, le mécanisme qui génère les

distances de la matrice $\mathbf{X}$ est indépendant du mécanisme qui génère les distances de la matrice $\mathbf{Y}$ (Mantel,1967). Cela revient au cas où les distances de deux ensembles seraient calculées à partir de deux ensembles indépendants d'individus (Manly,1991). L'hypothèse alternative est souvent unilatérale, dans le sens d'une relation positive ou négative entre $\mathbf{X}$ et $\mathbf{Y}$.

La statistique du test est le produit croisé des éléments des deux matrices :

$$\mathbf{T_{XY}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{X}_{ij} \mathbf{Y}_{ij}.$$

Puisqu'on considère des distances symétriques et que la distance entre un individu et lui même est nulle, la statistique $\mathbf{T_{XY}}$ est équivalente à la statistique $\mathbf{U_{XY}}$ définie par :

$$\mathbf{U_{XY}} = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathbf{X}_{ij} \mathbf{Y}_{ij}.$$

La distribution sous l'hypothèse nulle de $\mathbf{U_{XY}}$ peut être obtenue par l'approche de permutation que Mantel (1967) a nommé «approche de population finie». On a n positions des individus dans l'espace et n positions des individus dans le temps. L'hypothèse H_0 «no clustering» est équivalente à celle que les emplacements des cas dans l'espace sont regroupés aléatoirement avec les emplacements des cas dans le temps. Il y a au total $n!$ ensembles correspondants équiprobables.

Comme aux chapitres précédents, soit $\pi = (\pi_1, \cdots, \pi_n)$ une permutation des entiers $(1, \cdots, n)$ et soit Π l'ensemble de toutes les $n!$ permutations possibles. La loi de la statistique $\mathbf{U_{XY}}$ sous l'hypothèse nulle s'obtient en donnant une probabilité $1/n!$ pour les $n!$ valeurs $\mathbf{U_{XY}}$ correspondant à $\pi \in \Pi$. Chaque valeur de $\mathbf{U_{XY}}$ correspond à une permutation des n individus sur lesquels on a mesuré les deux (groupes de) variables, par exemple temps et espace.

Notons ici que ce sont les individus qui sont les unités de permutation et non pas les distances. L'hypothèse nulle H_0 affirme que le vecteur de variables obtenu pour un individu peut être observé pour tout autre individu. Cela veut dire que, si H_0 est

vraie, alors on peut permuter les individus, recalculer les matrices de distances après chaque permutation et finalement calculer la statistique du test. Ainsi la condition d'échangeabilité des observations nécessaire pour effectuer un test de permutation est satisfaite.

Pour faire le test de permutation, la stratégie de permutation est similaire à celle décrite dans le premier chapitre. On permute les individus dans l'une des matrices de données brutes et on recalcule les distances à partir des données permutées. Finalement, on recalcule la statistique du test.

$$\mathbf{U}_{\mathbf{X}\mathbf{Y}^\pi} = \sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathbf{X}_{ij}\mathbf{Y}^{\pi}_{ij}, \qquad \pi \in \Pi. \tag{4.1}$$

De fait, il y a un algorithme qui donne les mêmes résultats en permutant les lignes et leurs colonnes correspondantes dans la matrice des distances (voir section 4.3).

Si le nombre d'individus n'est pas petit, le calcul direct de la loi de la statistique sous H_0 devient «impraticable» à cause du grand nombre de permutations possibles. Alors, en pratique, on utilise des approches de calcul efficaces comme la méthode de Monte Carlo.

Ainsi, on utilise la distribution empirique de $\mathbf{U}_{\mathbf{X}\mathbf{Y}}$ pour calculer la probabilité de trouver des valeurs de $\mathbf{U}_{\mathbf{X}\mathbf{Y}}$ aussi extrêmes que la valeur observée $\mathbf{U}_{\mathbf{X}\mathbf{Y}_{obs}}$ par chance seulement. L'hypothèse alternative considère souvent une association positive entre les éléments des deux matrices.

La probabilité critique (p-valeur) pour tester la non existence d'un regroupement «clustering» est :

$$P - \text{valeur} = Pr(\mathbf{U}_{\mathbf{X}\mathbf{Y}^\pi} > \mathbf{U}_{\mathbf{X}\mathbf{Y}_{obs}}).$$

Mantel a présenté aussi une approximation asymptotique par la loi normale de la statistique $\mathbf{U}_{\mathbf{X}\mathbf{Y}}$, et a calculé son espérance et sa variance. Cependant, Mielke (1978) a démontré que $\mathbf{U}_{\mathbf{X}\mathbf{Y}}$ n'est pas toujours asymptotiquement normale.

4.1.2 Le test de Mantel modifié

Le critère $\mathbf{U_{XY}}$ proposé par Mantel est une mesure peu familière dont l'échelle varie d'un problème à un autre. Pour cela, Smoose et al. (1986) ont proposé de standardiser ce critère. En effet, $\mathbf{U_{XY}}$ est le produit croisé des éléments des deux matrices de distances $\mathbf{X}_{ij}$ et $\mathbf{Y}_{ij}$. On a déjà remarqué qu'il existe $N = n(n-1)/2$ éléments différents dans les matrices $\mathbf{X}$ et $\mathbf{Y}$, car $\mathbf{X}_{ii} = \mathbf{Y}_{ii} = 0$ et $\mathbf{X}_{ij} = \mathbf{X}_{ji}$ et $\mathbf{Y}_{ij} = \mathbf{Y}_{ji}$. On calcule les moyennes des éléments des deux matrices $\overline{\mathbf{X}} = \sum_{i<j}(\mathbf{X}_{ij}/N)$ et $\overline{\mathbf{Y}} = \sum_{i<j}(\mathbf{Y}_{ij}/N)$.

La somme corrigée d'un produit croisé est calculée par :

$$SP(\mathbf{X}, \mathbf{Y}) = \mathbf{U_{XY}} - N\overline{\mathbf{X}}\,\overline{\mathbf{Y}}. \tag{4.2}$$

En correspondance avec cette somme de produits corrigée on a aussi une paire des sommes au carré corrigées pour les éléments de chacune des matrices :

$$SS(\mathbf{X}) = \sum_{i,j}(\mathbf{X}_{ij} - \overline{\mathbf{X}})^2; \tag{4.3}$$

$$SS(\mathbf{Y}) = \sum_{i,j}(\mathbf{Y}_{ij} - \overline{\mathbf{Y}})^2. \tag{4.4}$$

Notons que $SP(\mathbf{X}, \mathbf{Y})$ change quand les éléments d'une des matrices sont permutés, alors que les sommes $SS(\mathbf{X})$ et $SS(\mathbf{Y})$ restent invariantes. La corrélation entre les éléments des deux matrices est égale à :

$$\mathbf{r_{XY}} = \frac{SP(\mathbf{X}, \mathbf{Y})}{(SS(\mathbf{X})SS(\mathbf{Y}))^{\frac{1}{2}}} = \frac{\mathbf{U_{XY}} - N\overline{\mathbf{X}}\,\overline{\mathbf{Y}}}{(SS(\mathbf{X})SS(\mathbf{Y}))^{\frac{1}{2}}}.$$

Ainsi, les statistiques $\mathbf{U_{XY}}$ et $\mathbf{r_{XY}}$ sont équivalentes et restent en relation monotone. Cela montre que la procédure de Mantel est vraiment une analyse de régression pour des distances. L'avantage d'utiliser $\mathbf{r}^2_{\mathbf{XY}}$ au lieu de $\mathbf{U_{XY}}$ est que $\mathbf{r}^2_{\mathbf{XY}}$ varie entre 0 et 1, donc l'association entre les deux matrices est facile à interpréter. Par contre, l'utilisation de $\mathbf{r}^2_{\mathbf{XY}}$ ou $\mathbf{U_{XY}}$ ne change rien à la probabilité critique du test.

Selon Smoose et al. (1986) le manque d'indépendance entre les distances n'est pas un problème, puisqu'on trouve la distribution de la statistique du test par la méthode de

permutation. Cela résulte du fait que les mesures des individus sont échangeables sous H_0. Par contre, ici on démontre que sous certaines suppositions les distances restent échangeables sous H_0, même si elles sont dépendantes (voir section 4.3).

4.2 Le test de Mantel partiel

Il y a beaucoup d'applications où on peut s'intéresser à utiliser plus de deux matrices. Soit un tel vecteur de matrices $(\mathbf{X}_1, \cdots, \mathbf{X}_p), p > 2$ explicatives pour prédire les éléments de la matrice de réponse $\mathbf{Y}$. Plusieurs auteurs ont présenté une généralisation du test de Mantel simple pour de telles situations (Smouse et al. 1986 ; Manly 1997). Dans un tel contexte, il est souvent le cas que les éléments des différentes matrices $\mathbf{X}_i$ soient corrélés entre eux. Ainsi, il y a une certaine redondance de l'information. On a besoin de pouvoir évaluer comment les matrices individuelles prédisent la matrice de réponse $\mathbf{Y}$, mais on veut connaître l'information additionnelle d'une matrice particulière $\mathbf{X}_i$ en sachant que d'autres matrices ont été déjà dans le modèle.

On considère ici deux matrices explicatives $\mathbf{X}$ et $\mathbf{Z}$. La généralisation à plus de deux matrices est évidente. En remplaçant dans le modèle linéaire les variables (Y_i, X_i, Z_i) par les éléments des matrices $\mathbf{Y}$, $\mathbf{X}$ et $\mathbf{Z}$ on peut considérer le modèle

$$\mathbf{Y}_{ij} = \beta_1 \mathbf{X}_{ij} + \beta_2 \mathbf{Z}_{ij} + \epsilon_{ij}, i = 1, \cdots, n, j = 1, \cdots, n, \tag{4.5}$$

où ϵ_{ij} est une erreur aléatoire. Cela nous ramène à l'approche du chapitre II. Le seul aspect difficile est le test de signification des coefficients de régression (par exemple $H_0 : \beta_1 = 0$). Smoose et al. (1986) suggèrent que la procédure de permutation diffère si on considère que les matrices $\mathbf{X}$ et $\mathbf{Z}$ prédisent $\mathbf{Y}$, ou sont seulement corrélées avec $\mathbf{Y}$. On sait que les matrices $\mathbf{X}$ et $\mathbf{Z}$ sont corrélées, et on veut une procédure de permutation qui conserve cette dépendance à travers les permutations.

Smoose et al. (1986) ont proposé deux méthodes de permutation. La première, reprise aussi par Manly (1997), consiste à permuter la matrice dépendante $\mathbf{Y}$ en gardant les matrices explicatives $\mathbf{X}$ et $\mathbf{Z}$ fixes, et calculer la statistique du test désirée pour chaque permutation de $\mathbf{Y}$.

Smoose et al. (1986) notent que, si $\mathbf{X}$ et $\mathbf{Z}$ sont elles mêmes variables, possiblement mesurées avec erreur plutôt que fixes et associées à $\mathbf{Y}$, il est préférable de permuter une des matrices résiduelles plutôt que la matrice $\mathbf{Y}$. Pour cela, on fait la régression de chacune des matrices de distances $\mathbf{Y}$ et $\mathbf{X}$ sur la matrice $\mathbf{Z}$ pour obtenir deux matrices résiduelles $\mathbf{R_{YZ}}$ et $\mathbf{R_{XZ}}$. La corrélation entre les éléments correspondants des deux matrices résiduelles représente la corrélation partielle entre les éléments des matrices de distances $\mathbf{X}$ et $\mathbf{Y}$ en tenant compte de $\mathbf{Z}$. Pour faire le test de permutation, on permute la matrice residuelle $\mathbf{R_{YZ}}$ pour obtenir une matrice $\mathbf{R^{\pi}_{YZ}}$ et on calcule la corrélation entre les matrices $\mathbf{R^{\pi}_{YZ}}$ et $\mathbf{R_{XZ}}$ pour chaque permutation. Notons que c'est la même technique de permutation que Kennedy (1995) a proposé (voir chapitre II, section 2.3.3).

Dans nos applications numériques, nous proposons de généraliser la technique proposée par Freedman et Lane (1983) à des matrices de distances. La description est la même que pour le cas de la régression habituelle. Dans ce cas, les éléments des matrices $\mathbf{Y}$ et $\mathbf{X}$ sont régressés sur les éléments de la matrice $\mathbf{Z}$, pour obtenir deux matrices de résidus $\mathbf{R_{XZ}}$ et $\mathbf{R_{YZ}}$. Après, on procède de la façon suivante :

- on permute la matrice des résidus $\mathbf{R_{YZ}}$ pour obtenir une matrice $\mathbf{R^{\pi}_{YZ}}$;
- on ajoute la matrice permutée au modèle de régression pour construire une nouvelle matrice dépendante $\mathbf{Y}_{\pi}$, où $\mathbf{Y}_{\pi} = \widehat{\beta}\mathbf{Z} + \mathbf{R^{\pi}_{YZ}}$;
- cette matrice $\mathbf{Y}_{\pi}$ est régressée de nouveau sur la matrice Z pour trouver une nouvelle matrice résiduelle $\mathbf{R_{Y_{\pi}Z}}$;
- la statistique du test est la corrélation entre les éléments des deux matrices résiduelles $\mathbf{R_{Y_{\pi}Z}}$ et $\mathbf{R_{XZ}}$.

À la section 4.4 nous présentons une application de ces tests.

4.3 Deux résultats théoriques

4.3.1 Équivalence entre deux façons de permuter

On va montrer l'équivalence entre la permutation des lignes de la matrice de données brutes, $\mathbf{A}$, et la permutation des lignes avec leurs colonnes correspondantes dans la matrice de distances calculée à partir de ces données.

Supposons qu'on a n individus et qu'on a mesuré p variables pour chacun d'entre eux pour obtenir la matrice de données brutes $\mathbf{A}_{n\times p}$. On calcule la matrice de distances $\mathrm{D}_{n\times n}$ entre les mesures sur les individus. Alors, on a : $d_{ij} = d_{ji}$ et $d_{ii} = 0$, $i = 1, \cdots, n$, et $j = 1, \cdots, n$.

Soit, par exemple, une matrice $\mathrm{D}_{4\times 4}$:

$$\mathbf{D} = \begin{pmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{21} & 0 & d_{23} & d_{24} \\ d_{31} & d_{32} & 0 & d_{34} \\ d_{41} & d_{42} & d_{43} & 0 \end{pmatrix}.$$

Regardons dans cet exemple l'effet de la permutation de deux lignes dans la matrice de données brutes, $\mathbf{A}$, sur la matrice de distances, $\mathbf{D}$. Si on permute les lignes i et j, on obtient une matrice $\widetilde{\mathbf{A}}$ à laquelle correspond une matrice de distances $\widetilde{\mathbf{D}}$ telle que $\widetilde{d}_{ik}$ est la distance entre l'individu k et l'individu i de la matrice $\widetilde{\mathbf{A}}$ qui correspond à la distance entre les les individus k et j de la matrice $\mathbf{A}$. Alors $\widetilde{d}_{ik} = d_{jk} = d_{kj} = \widetilde{d}_{ki}$ et $\widetilde{d}_{jk} = d_{ik} = d_{ki} = \widetilde{d}_{kj}$.

Supposons que, dans notre exemple, on permute la ligne 2 avec la ligne 3 dans la matrice $\mathbf{A}$. Alors, par le raisonnement qu'on vient de faire, la matrice $\widetilde{\mathbf{D}}$ suivante :

$$\widetilde{\mathbf{D}} = \begin{pmatrix} 0 & \widetilde{d}_{12} = d_{13} & \widetilde{d}_{13} = d_{12} & \widetilde{d}_{14} = d_{14} \\ \widetilde{d}_{21} = d_{31} & 0 & \widetilde{d}_{23} = d_{32} & \widetilde{d}_{24} = d_{34} \\ \widetilde{d}_{31} = d_{21} & \widetilde{d}_{23} = d_{23} & 0 & \widetilde{d}_{34} = d_{24} \\ \widetilde{d}_{41} = d_{41} & \widetilde{d}_{42} = d_{43} & \widetilde{d}_{43} = d_{42} & 0 \end{pmatrix}$$

est égale à

$$\begin{pmatrix} 0 & d_{13} & d_{12} & d_{14} \\ d_{31} & 0 & d_{32} & d_{34} \\ d_{21} & d_{23} & 0 & d_{24} \\ d_{41} & d_{43} & d_{42} & 0 \end{pmatrix}.$$

Donc, on peut voir que la matrice $\widetilde{\mathbf{D}}$ s'obtient de la façon suivante à partir de la matrice $\mathbf{D}$: on permute la ligne 2 avec la ligne 3 en même temps qu'on permute la colonne 2 avec la colonne 3. Ainsi, permuter deux lignes dans la matrice de données brutes revient à permuter les lignes en même temps que leurs colonnes correspondantes dans la matrice de distances. Donc, dans les tests de permutation, pour plus d'efficacité, et pour éviter des calculs inutiles, on peut effectuer les permutations directement dans la matrice de distances.

4.3.2 Échangeabilité des distances

On commence par démontrer ici que, si les mesures sur les individus sont indépendantes ou échangeables, les distances entre ces mesures restent échangeables. Cela se base sur le théorème qui suit.

Théorème 4.1 *(Randles and Wolfe 1979, théorème 1.3.7)*

Si les vecteurs $\boldsymbol{X}$ *et* $\boldsymbol{Y}$ *sont de même loi,* $\boldsymbol{X} \underline{\underline{d}} \ \boldsymbol{Y}$*, et* $\boldsymbol{U}(.)$ *est une fonction mesurable définie sur le support commun des variables, alors :*

$$\boldsymbol{U}(\boldsymbol{X}) \ \underline{\underline{d}} \ \boldsymbol{U}(\boldsymbol{Y}).$$

■

Notons par $\underset{\sim}{\mathbf{V}}_i = (V_{i1}, \cdots, V_{ip})$ le vecteur des p variables pour l'individu i. La matrice de données brutes est $\mathbf{V} = (\underset{\sim}{\mathbf{V}}_1, \underset{\sim}{\mathbf{V}}_2, \cdots, \underset{\sim}{\mathbf{V}}_n)$. Puisque les vecteurs $\underset{\sim}{\mathbf{V}}_i$ sont supposés indépendants ou échangeables, alors, pour une permutation $\pi \in \Pi$, on a :

$$(\underset{\sim}{\mathbf{V}}_1, \underset{\sim}{\mathbf{V}}_2, \cdots, \underset{\sim}{\mathbf{V}}_n) \ \underline{\underline{d}} \ (\underset{\sim}{\mathbf{V}}_{\pi_1}, \underset{\sim}{\mathbf{V}}_{\pi_2}, \cdots, \underset{\sim}{\mathbf{V}}_{\pi_n}).$$

Soit la fonction $\Delta(.)$ une distance entre les mesures de deux individus i et j. Ainsi, on prend :

$$\mathbf{U}(\underset{\sim}{\mathbf{V}}_1, \underset{\sim}{\mathbf{V}}_2, \cdots, \underset{\sim}{\mathbf{V}}_n) = \Delta(\underset{\sim}{\mathbf{V}}_i, \underset{\sim}{\mathbf{V}}_j),$$

et

$$\mathbf{U}(\underset{\sim}{\mathbf{V}}_{\pi_1}, \underset{\sim}{\mathbf{V}}_{\pi_2}, \cdots, \underset{\sim}{\mathbf{V}}_{\pi_n}) = \Delta(\underset{\sim}{\mathbf{V}}_{\pi_i}, \underset{\sim}{\mathbf{V}}_{\pi_j}).$$

On applique le théorème précédent à la fonction $\mathbf{U}(.) = \Delta(.)$ et on obtient, par conséquent, que les distances restent échangeables même si elle ne sont pas indépendantes, d'où

$$\Delta(\underset{\sim}{\mathbf{V}}_i, \underset{\sim}{\mathbf{V}}_j) \underset{=}{d} \Delta(\underset{\sim}{\mathbf{V}}_{\pi_i}, \underset{\sim}{\mathbf{V}}_{\pi_j}).$$

Donc, on peut considérer les distances comme des variables doublement indicées et on peut appliquer directement les méthodes présentées dans le chapitre II. Cela permet de formuler le test de Mantel de la façon suivante. Soit deux groupes de variables $\mathbf{V} = (V^{(1)}, \cdots, V^{(p)})$ et $\mathbf{W} = (W^{(1)}, \cdots, W^{(q)})$ qu'on mesure sur n individus. À ces groupes de variables, on associe les distances respectives $\Delta_{ij}^{(V)}$ et $\Delta_{ij}^{(W)}$, $i, j = 1, \cdots, n$.

L'hypothèse nulle est qu'il n'y a pas de relation entre les deux groupes de variables pour le même individu, et donc $\underset{\sim}{\mathbf{V}}_i$ est indépendant de $\underset{\sim}{\mathbf{W}}_i$, $i = 1, \cdots, n$. De plus, les individus sont "indépendants" et donc $\underset{\sim}{\mathbf{V}}_i$ est indépendant de $\underset{\sim}{\mathbf{V}}_j$, $i \neq j$. On conclut ainsi que $\underset{\sim}{\mathbf{V}}_i$ est indépendant de $\underset{\sim}{\mathbf{W}}_j$, $i \neq j$.

Finalement, on a que $\Delta_{ij}^{(V)} = f(\underset{\sim}{\mathbf{V}}_i, \underset{\sim}{\mathbf{V}}_j)$ est indépendant de $\Delta_{ij}^{(W)} = f(\underset{\sim}{\mathbf{W}}_i, \underset{\sim}{\mathbf{W}}_j)$ pour i et j fixés. De plus, les vecteurs de distances $\Delta_{ij}^{(V)} | i = 1, \cdots, n; j \leq i$ et $\Delta_{ij}^{(W)} | i = 1, \cdots, n; j \leq i$ sont des vecteurs de variables échangeables.

4.4 Exemple d'application

On illustre la méthodologie sur un exemple de données du chapitre IX de Manly (1997), où on étudie la relation entre les conditions environnementales et les variations génétiques pour 21 colonies de papillons «Euphydryas Editha», en tenant compte des distances géographiques entre ces colonies. Ces donnéees proviennent d'une étude faite par McKechnie et al. (1975) en Californie et Oregon. On commence par faire le test de Mantel pour évaluer les relations entre les trois matrices de distances prises deux à deux. Puis on applique le test de Mantel partiel en utilisant les différentes techniques présentées au chapitre II pour évaluer s'il y a une relation entre les distances génétiques et les distances environnementales, étant donné les distances géographiques.

On note par $\mathbf{G}$ la matrice de distances génétiques, par $\mathbf{E}$ la matrice de distances environnementales et par $\mathbf{S}$ la matrice de distances géographiques.

Les éléments de la matrice $\mathbf{E}$ sont obtenus par le calcul de la distance euclidienne standardisée basée sur dix variables qui décrivent l'environnement (l'altitude, les précipitations annuelles et huit variables de température). Notons par $\mathbf{E}^*_{ik}$ la valeur de la $k^{ème}$ variable environnementale pour la $i^{ème}$ colonie après standardisation. Ainsi, la moyenne est 0 et l'écart type est 1. La distance environnementale entre la colonie i et la colonie j est décrite par l'équation suivante :

$$\mathbf{E}_{ij} = \sqrt{\sum_{k=1}^{10} (\mathbf{E}^*_{ik} - \mathbf{E}^*_{jk})^2}. \tag{4.6}$$

Les éléments de la matrice des distances génétiques sont obtenus on utilisant la matrice des différences génétiques de l'enzyme hexokinase «Hexokinase» entre deux colonies, cette matrice est décrite par l'équation suivante :

$$\mathbf{G}_{ij} = \sum_{r=1}^{3} |p_{ir} - p_{jr}|. \tag{4.7}$$

où p_{ir} représente la fréquence de l'allèle r dans la $i^{ème}$ colonie. Notons qu'il y a trois allèles responsables de l'enzyme hexokinase.

Le calcul de la distance géographique entre deux colonies se fait en utilisant la distance en ligne droite entre les colonies avec une unité d'approximativement 111.2 km.

Les trois matrices de distances environnementales, génétiques et géographiques de ces données sont présentées dans les tableaux suivants :

Tableau 4.1 la matrice de distances environnementales

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1.05																			
3	2.26	1.40																		
4	1.78	0.87	0.55																	
5	1.78	0.87	0.55	0.01																
6	1.78	0.87	0.55	0.01	0.00															
7	2.26	1.30	0.45	0.51	0.50	0.50														
8	2.09	1.17	0.33	0.37	0.38	0.38	0.40													
9	1.95	1.02	0.46	0.19	0.19	0.19	0.34	0.31												
10	2.40	1.41	0.57	0.68	0.68	0.68	0.29	0.46	0.53											
11	2.40	1.41	0.57	0.68	0.68	0.68	0.27	0.47	0.53	0.03										
12	2.09	1.16	0.82	0.73	0.74	0.74	0.82	0.55	0.76	0.71	0.74									
13	2.16	1.32	0.73	0.79	0.80	0.80	0.90	0.53	0.81	0.81	0.84	0.38								
14	1.09	1.08	1.89	1.54	1.55	1.55	1.99	1.67	1.70	2.03	2.05	1.49	1.48							
15	2.63	2.32	2.60	2.46	2.47	2.47	2.72	2.37	2.57	2.60	2.63	1.93	1.91	1.62						
16	1.85	1.29	1.69	1.45	1.46	1.46	1.74	1.42	1.56	1.66	1.68	0.99	1.05	0.93	1.09					
17	2.11	1.28	1.24	1.11	1.12	1.12	1.24	0.96	1.15	1.09	1.12	0.49	0.60	1.35	1.58	0.65				
18	4.03	3.91	4.29	4.16	4.17	4.17	4.40	4.07	4.26	4.27	4.29	3.61	3.61	3.15	1.72	2.76	3.24			
19	4.72	4.62	4.85	4.78	4.79	4.79	5.01	4.66	4.88	4.87	4.90	4.22	4.17	3.78	2.34	3.42	3.87	0.93		
20	5.49	5.53	5.93	5.80	5.82	5.82	6.07	5.72	5.91	5.94	5.97	5.28	5.25	4.67	3.37	4.41	4.92	1.72	1.25	
21	6.53	6.53	6.88	6.78	6.79	6.79	7.01	6.68	6.88	6.87	6.90	6.22	6.20	5.70	4.33	5.39	5.86	2.65	2.09	1.06

Tableau 4.2 la matrice des distances génétiques

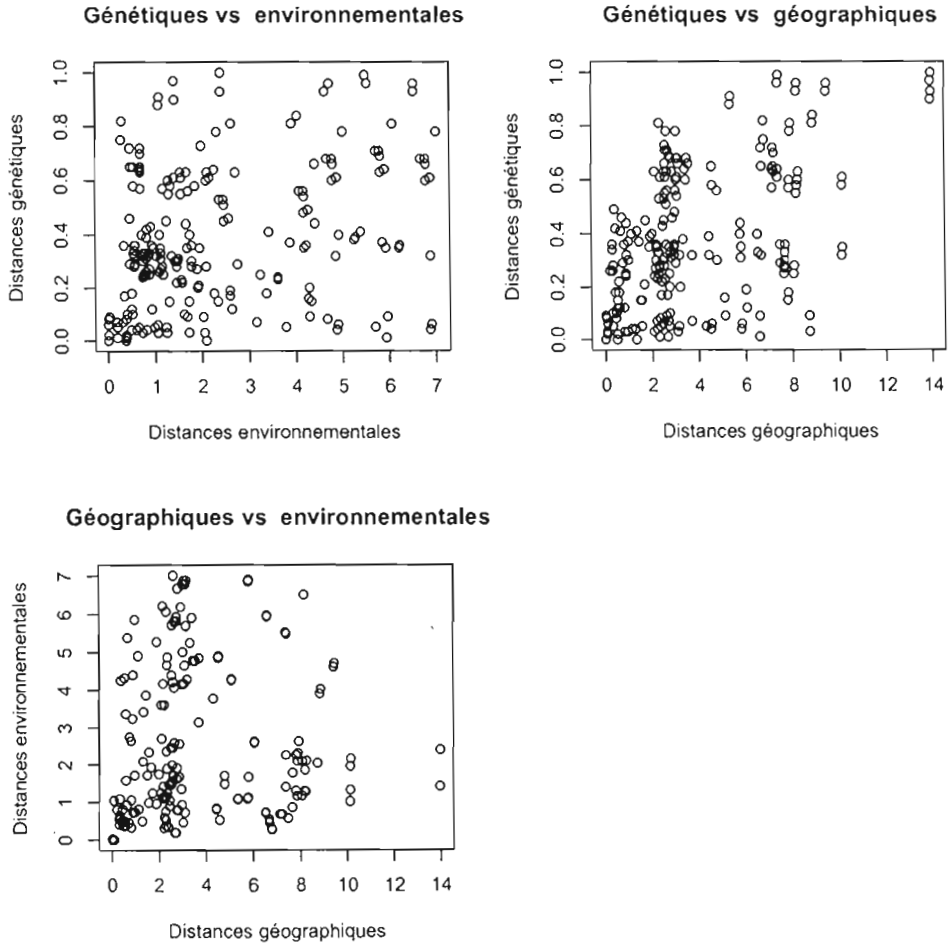
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0.03																			
3	0.64	0.61																		
4	0.28	0.25	0.36																	
5	0.30	0.27	0.34	0.02																
6	0.36	0.33	0.28	0.08	0.06															
7	0.18	0.15	0.46	0.10	0.12	0.18														
8	0.28	0.25	0.36	0.00	0.02	0.08	0.10													
9	0.35	0.32	0.29	0.07	0.05	0.01	0.17	0.07												
10	1.00	0.97	0.36	0.72	0.70	0.64	0.82	0.72	0.65											
11	0.93	0.90	0.29	0.65	0.63	0.57	0.75	0.65	0.58	0.09										
12	0.60	0.57	0.04	0.32	0.30	0.24	0.42	0.32	0.26	0.39	0.33									
13	0.61	0.58	0.03	0.33	0.31	0.25	0.43	0.33	0.26	0.39	0.32	0.01								
14	0.91	0.88	0.27	0.63	0.61	0.55	0.73	0.63	0.56	0.09	0.32	0.31	0.30							
15	0.81	0.78	0.17	0.53	0.51	0.45	0.63	0.53	0.46	0.19	0.03	0.21	0.20	0.10						
16	0.58	0.55	0.09	0.30	0.28	0.22	0.40	0.30	0.23	0.44	0.12	0.05	0.06	0.36	0.26					
17	0.63	0.60	0.05	0.35	0.33	0.27	0.45	0.35	0.28	0.40	0.35	0.04	0.04	0.32	0.22	0.05				
18	0.84	0.81	0.20	0.56	0.54	0.48	0.66	0.56	0.49	0.16	0.31	0.24	0.23	0.07	0.03	0.29	0.25			
19	0.96	0.93	0.32	0.68	0.66	0.60	0.78	0.68	0.61	0.04	0.09	0.36	0.35	0.05	0.15	0.41	0.37	0.12		
20	0.99	0.96	0.35	0.71	0.69	0.63	0.81	0.71	0.64	0.01	0.09	0.39	0.38	0.08	0.18	0.44	0.40	0.15	0.03	
21	0.96	0.93	0.32	0.68	0.66	0.60	0.78	0.68	0.61	0.04	0.06	0.36	0.35	0.05	0.15	0.41	0.37	0.12	0.00	0.01

Tableau 4.3 la matrice des distances géographiques

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0.07																			
3	7.37	7.36																		
4	7.64	7.63	0.30																	
5	7.66	7.65	0.30	0.08																
6	7.68	7.67	0.34	0.04	0.10															
7	7.81	7.79	0.71	0.47	0.53	0.43														
8	8.07	8.05	0.80	0.50	0.53	0.46	0.31													
9	10.13	10.11	3.01	2.70	2.72	2.66	2.39	2.20												
10	13.97	13.94	7.46	7.16	7.20	7.12	6.77	6.66	4.56											
11	13.98	13.95	7.47	7.18	7.21	7.14	6.79	6.68	4.57	0.01										
12	7.86	7.84	1.11	0.90	0.97	0.87	0.44	0.66	2.27	6.50	6.52									
13	10.14	10.12	3.09	2.78	2.81	2.74	2.45	2.28	0.21	4.43	4.44	2.29								
14	5.36	5.34	2.33	2.53	2.57	2.55	2.55	2.85	4.78	8.71	8.73	2.53	4.78							
15	7.93	7.89	2.65	2.52	2.60	2.50	2.10	2.31	2.87	6.04	6.06	1.66	2.76	2.75						
16	8.19	8.16	2.60	2.44	2.51	2.41	1.99	2.17	2.55	5.78	5.80	1.55	2.44	2.96	0.32					
17	8.25	8.22	2.35	2.17	2.24	2.14	1.71	1.87	2.30	5.76	5.78	1.27	2.21	2.95	0.58	0.32				
18	8.88	8.84	3.20	3.00	3.07	2.97	2.54	2.65	2.39	5.09	5.11	2.10	2.23	3.69	0.95	0.74	0.87			
19	9.46	9.43	3.71	3.49	3.55	3.45	3.02	3.09	2.36	4.51	4.53	2.60	2.18	4.30	1.56	0.34	1.44	0.61		
20	7.40	7.36	2.74	2.67	2.74	2.65	2.30	2.56	3.41	6.58	6.59	1.90	3.32	2.34	0.58	0.90	1.11	1.49	2.07	
21	8.18	8.15	3.18	3.04	3.12	3.02	2.61	2.80	3.07	5.81	5.83	2.17	2.94	3.15	0.53	0.65	0.97	0.80	1.31	0.81

La figure 4.1 suivante présente les distances tracées l'une contre l'autre. D'après cette figure, la relation entre les distances environnementales et les distances génétiques paraît faible. En outre, elle indique une relation possible entre les distances génétiques et les distances géographiques d'une part et les distances environnementales d'autre part.

Figure 4.1 Les graphiques des matrices de distances l'une versus l'autre



Pour tester l'association entre les matrices deux à deux, on a utilisé le test de Mantel. L'argument pour faire le test de Mantel est de considérer les n colonies de papillons pour lesquelles les distances sont calculées comme un échantillon aléatoire d'une plus grande population de colonies potentielles qui pourraient être étudiées.

Le test de Mantel a été fait avec 4999 permutations et on a utilisé la statistique **U** décrite à la formule (4.1) . Le tableau (4.4) présente les résultats obtenus :

Tableau 4.4 Résultats du test de Mantel simple

	Corrélation	P-valeur
Génétiques vs Géographiques	0.48534	0.0002
Génétiques vs Environnementales	0.2915438	0.0072
Environnementales vs Géographiques	0.02640319	0.3894

On voit bien que les distances génétiques sont significativement corrélées aux distances géographiques et environnementales. Par contre, le test de Mantel n'est pas significatif pour la corrélation entre les distances géographiques et les distances environnementales, ce qui est surprenant à premier abord. D'autres part, selon la carte géographique, les données qui proviennent de la même latitude (donc avec les mêmes conditions environnementales, le même climat) ne sont pas nécessairement proches du point de vue géographique. Ces résultats sont consistants avec les représentations graphiques de la figure (4.1), à part la relation significative entre les distances génétiques et environnementales, qui est vraisemblablement due au grand nombre de colonies où autant les distances génétiques, que les distances environnementales, sont proches de zéro.

La relation significative entre les distances génétiques et les distances environnementales peut s'expliquer par l'adaptation au climat. La relation significative entre les distances génétiques et les distances géographiques peut s'expliquer par la migration entre les colonies proches.

On peut alors essayer d'expliquer les distances génétiques en fonction des distances

géographiques et les distances environnementales prises ensemble. On se retrouve ici avec un modèle de régression avec **E** et **S** fixes qui explique les distances génétiques $\mathbf{G}_{n\times n}$ par les distances géographiques $\mathbf{S}_{n\times n}$ et environnementales $\mathbf{E}_{n\times n}$. Ce modèle est de la forme

$$\mathbf{G}_{ij} = \beta_0 + \beta_1\mathbf{E}_{ij} + \beta_2\mathbf{S}_{ij} + \epsilon_{ij}, i = 1, \cdots, n \, et \, j = 1, \cdots, n\,. \tag{4.8}$$

où β_1 mesure la relation entre $\mathbf{G}_{ij}$ et $\mathbf{E}_{ij}$ en tenant compte l'effet de $\mathbf{S}_{ij}$ et β_2 mesure la relation entre $\mathbf{G}_{ij}$ et $\mathbf{S}_{ij}$ en tenant compte l'effet de $\mathbf{E}_{ij}$, tandis que ϵ_{ij} sont des erreurs aléatoires indépendantes.

Après, on s'intéresse à tester l'hypothèse nulle $H_0 : \beta_1 = 0$. C'est à dire on veut savoir si relation entre les distances génétiques entre les colonies de papillons et leurs distances environnementales est significative lorsqu'on tient compte de leurs distances géographiques.

On a régressé les distances génétiques sur les distances géographiques et les distances environnementales et on trouvé l'équation de régression ajustée suivante :

$$\widehat{\mathbf{G}}_{ij} = 0.133 + 0.041\mathbf{E}_{ij} + 0.037\mathbf{S}_{ij}\,. \tag{4.9}$$

On peut utiliser la régression usuelle pour ajuster l'équation précédente, mais on ne peut pas appliquer les tests usuels parce que, si on utilise les distances, les hypotèses du modèle linéaire ne sont pas satisfaites (les distances ne sont pas indépendantes). Cela suggère l'utilisation des différentes techniques de permutation présentées dans le chapitre précèdent, pour évaluer la signification des coefficients de régression partiels.

Afin d'effectuer nos tests on a adapté les techniques de permutation de Freedman et Lane (1983), Kennedy (1995) et Manly (1997) sur les matrices de distances. On a utilisé la corrélation partielle comme statistique de test et on a généré sa distribution approximative en utilisant 4999 permutations. Notons que la méthode de Freedman et

Lane (1983) n'a pas encore été utilisée dans le cas des matrices.

Pour la méthode de Manly (1997) on a gardé les deux matrices explicatives (ici **S** et **E**) fixes et on a permuté aléatoirement l'ordre des 21 colonies dans la matrice des distances génétiques, **G**. Pour la méthode de Kennedy (1995) on a permuté la matrice des résidus de la régression des distances génétiques sur les distances géographiques $\mathbf{R_{GS}}$ et on a utilisé la corrélation entre les matrices des résidus $\mathbf{R^{\pi}_{GS}}$ et $\mathbf{R_{ES}}$ (cela revient à appliquer le test de Mantel simple sur les matrices des résidus).

Pour la méthode de Freedman et Lane (1983), on a permuté la matrice des résidus de la régression des distances génétiques sur les distances géographiques et on a utilisé cette matrice $\mathbf{R^{\pi}_{GS}}$ pour construire une nouvelle matrice de distances génétiques $\mathbf{G}_{\pi}$. $\mathbf{G}_{\pi}$ est régressée de nouveau sur **S** et la statistique du test est la corrélation entre la matrice $\mathbf{R_{G_{\pi}S}}$ et $\mathbf{R_{ES}}$.

Le programme décrit dans l'annexe a donné une corrélation partielle égale à $\mathbf{R}^2_{GE.S} = 0.1017002$ et la même probabilité critique ou (p-valeur) de 0.000242 pour les trois méthodes. Cela est peut-être dû au fait qu'on a un grand nombre d'observations $n = 210$, et on a déjà vu dans les simulations qu'à partir d'une certaine taille (environ 45) les méthodes donnent des résultats similaires.

Pour le test sur $H_0 : \beta_2 = 0$ on a trouvé une corrélation partielle entre les distances génétiques et les distances géographiques en tenant compte des distances environnementales de $R^2_{GS.E} = 0.2495090$. Le test est significatif avec probabilité critique (p-valeur) de 0.000237

CONCLUSION

Ce mémoire est centré sur l'application des techniques de permutation pour tester les coefficients de corrélation partiels dans la régression multiple. Nous avons vu que ce genre de méthodologie est difficile à réaliser puisque la permutation des valeurs de la variable dépendante Y n'est pas appropriée. Cela est dû au fait que, sous l'hypothèse nulle $H_0 : \beta_1 = 0$, les valeurs observées de Y ne sont pas échangeables. Ce sont plutôt les erreurs qui sont échangeables sous l'hypothèse nulle.

En effet, un test exact reposerait sur la connaissance du paramètre de la régression de Y sur Z, λ_Y, sous H_0. Cela permettrait de permuter les résidus de cette régression, $\epsilon_{YZ} = Y_i - \lambda_Y Z$. Évidemment, λ_Y est inconnu et aucun test de permutation exact n'est possible (Anderson et Robinson, 2001). Plusieurs auteurs ont proposé différentes façons pour construire un test de permutation approximatif (Freedman et Lane, 1983 ; Collins, 1987 ; Oja, 1987 ; Welch, 1990 ;Braak, 1992 ; Kennedy, 1995 ; Manly, 1997). Trois de ces tests (Freedman et Lane, 1983 ; Kennedy, 1995 ; Manly, 1997) sont présentés en détail dans ce travail. Ensuite, la présentation théorique est complétée par une comparaison par simulation entre ces trois façons différentes de faire un test de permutation approximatif. On a examiné l'effet de la taille de l'échantillon n, le degré de colinéarité entre les covariables, ρ_{XZ}, la valeur du coefficient de la covariable Z, β_2 et la loi de l'erreur aléatoire ϵ. Pour notre étude, on a introduit de nouvelles lois pour les erreurs par rapport à la littérature existante, par exemple le cas où les erreurs sont échangeables, mais pas indépendantes.

Les méthodes de Freedman et Lane (1983) et Kennedy (1995) se basent sur la permutation des résidus sous le modèle réduit, mais les façons de faire les permutations sont différentes. La méthode de Kennedy, utilise directement les résidus permutés dans le calcul de la statistique du test, et ainsi l'estimation du paramètre λ_Y reste fixe à

travers les permutations. La méthode de Freedman et Lane (1983) utilise les résidus permutés pour créer de nouvelles observations qui seront régressées de nouveau sur la covariable Z, ainsi, l'estimation du paramètre λ_Y change à travers des permutations. Manly (1997), quant à lui, a proposé de permuter Y en gardant Z et X fixes et, pour chaque permutation, utiliser les résidus de la régression des valeurs permutées de Y sur la covariable Z dans le calcul de la statistique du test.

Les résultats de nos simulations ont montré que la méthode de permutation proposée par Kennedy (1995) n'est pas équivalente à la méthode proposée par Freedman et Lane (1983). La méthode de Kennedy (1995) donne une erreur de type I gonflée, particulièrement lorsque la taille de l'échantillon est petite. Ces résultats concordent avec ce qui avait été trouvé dans la littérature. Pour ce qui est des trois autres méthodes (Manly, 1997 ; Freedman et Lane, 1983 et la méthode qu'on a proposée), elle ont donné des résultats équivalents dans la plupart des situations simulées.

Par contre, l'étude empirique d'Anderson et Legendre (1999) avait montré que même l'approche de Manly ne fonctionne pas toujours bien, car elle est sensible aux données extrêmes de la covariable Z, investigation qu'on n'a pas repris dans notre étude. On constate ainsi que la méthode proposée par Freedman et Lane (1983) est la plus proche du test de permutation exact conceptuel.

Une avenue intéressante à explorer serait le développement d'un test ressemblant au test de Kennedy (1996) mais qui incorporait une estimation de type *bootstrap*. Cette variante pourrait améliorer la performance de ce dernier test.

On remarque aussi que pour des tailles d'échantillon suffisamment grandes, $n \geq 45$, les quatre méthodes sont équivalentes au niveau de l'erreur type I et de la puissance. Cela veut dire qu'on pourrait choisir celle qui demande moins de calcul. Dans ce cas, on peut recommander d'utiliser la méthode de Kennedy (1995) puisqu'elle n'exige pas de refaire la régression après chaque permutation.

Finalement, dans un dernier chapitre de ce travail, on a présenté une généralisation des

tests de permutation pour les coefficients partiels de régression lorsque les variables sont remplacées par des éléments de matrices de distances. Cette généralisation est de plus en plus utilisée dans l'analyse des données multivariées (écologie, biologie, agriculture...) lorsque les données ne respectent pas, généralement, les suppositions nécessaires pour les tests paramétriques traditionnels. Nous avons proposé d'appliquer les trois différentes techniques de permutation sur un exemple de données génétiques. Il s'agit de l'étude de la relation des distances génétiques en fonction des distances géographiques et des distances environnementales. L'analyse a donné des résultats équivalents, mais on pense que cela est dû à la grande taille d'échantillon de cet exemple. Notons que l'adaptation de la méthode de Freedman et Lane (1983) dans ce contexte (matrices de distances) n'avait pas été encore faite et elle pourrait s'avérer utile pour des données qui présentent des valeurs extrêmes en Z.

APPENDICE A

SIMULATION DES ERREURS DE LOI NORMALE DÉPENDANTES

Pour générer un vecteur de loi normale multivariée :

$$\mathbf{X} = (X_1, \cdots, X_n) \sim N(\mu, \Sigma)$$

tel que X_i et X_j sont dépendants avec une corrélation $\rho_{ij} \neq 0$, on doit procéder par les étapes suivantes :

1. On génère n variables $(Z_1, \cdots, Z_n)$ de loi normale centrée réduite avec $Z_i \sim N(0,1)$ indépendantes (Pour cela on utilise la fonction rnorm) ;
2. On fait la décomposition orthogonale de la matrice de covariance. Pour cela, on calcule les valeurs propres $(\lambda_1, \cdots, \lambda_n)$ de Σ et la matrice orthgonale P constituée de ces vecteurs propres $V_1, \cdots, V_n$. La matrice $\Sigma^{\frac{1}{2}}$ est éagle à :

$$\Sigma^{\frac{1}{2}} = P\,\Lambda^{\frac{1}{2}}\,P^T \tag{A.1}$$

où $\Lambda^{\frac{1}{2}}$ est une matrice diagonale composée des racines carrés des valeurs propres $\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n}$

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix} \tag{A.2}$$

3. on calcule les valeurs des variables $X_1, \cdots, X_n$ en utilisant l'équation suivante :

$$\mathbf{X} = \Sigma^{\frac{1}{2}}\mathbf{Z} \sim N(0, \Sigma) \tag{A.3}$$

APPENDICE A

PROGRAMMES

```
###### ERREUR DE TYPE I PUISSANCE EMPIRIQUE #################

simul_puissance<-function(n,beta1,beta2,rho,loi=c("uniforme","normaldep"))
{

# ce programme calcule l'erreur de type 1 lorsque beta1=0 # et la
puissance lorsque beta1 différent de 0 # d'abord on détermine les
constantes # taille, nombre de simulations, # coefficient de
regression partiel, nombre de permutation. # le modele est Y=
beta_1*X + beta_2* Z + e #l'hypothese alternative est que le
coefficient partiel beta-1

nperm<-999 nsimul<-2000

##############la fonction normal.bivariee"################### # #
#Cette fonction génère un couple bivarié de loi normale
 normal.bivariee<-function(n,rho)
 {
sigma<-matrix(c(1,rho,rho,1),nrow=2)
result<-eigen(sigma,symmetric=T)
```

```
# on construit une matrice diagonale "val" telle que ses diagonales
# sont les racines carrées des valeurs propres; vec représente la
# matrice orthogonale des vecteurs propres

val<-diag(sqrt(result$values)) vec<-result$vectors
sigma1demi<-vec%*%val%*%t(vec)

# on génère deux variables normales centrées réduites
#indépendantes on calcule deux variables normales corrélées
#X=(X1,X2)  A partir de la formule X = sigma^1/2 Z

Z<-matrix(rnorm(2*n),nrow=2,ncol=n) X<-matrix(0,nrow=2,ncol=n)
X<- sigma1demi%*%Z
X }

################  fonction  multinormale.dep "###############

# la fonction "multinormale.dep" simule des erreurs normales
#corrélées dans cette fonction on suppose que la corrélation entre
# les variables  epsilon_i et epsilon_j

multinormale.dep<-function(n){ rho1<-0.85

# la matrice de covariance est une matrice symétrique appelée
#sigma où sigmaii=1 et sigmaij = rho

ma<-matrix(rho1,nrow=n,ncol=n) di<-diag(1-rho1,nrow=n,ncol=n)
sigma<-ma + di
```

```
# on doit calculer la matrice sigma^1/2 nécessaire pour  trouver #
# des normales corrélées;  on va calculer les valeurs et les
#vecteurs propres de sigma  on construit une matrice diagonale
#"val" tel que ses diagonales sont les racines carees des valeurs
#propres.

result<-eigen(sigma,symmetric=T) val<-diag(sqrt(result$values))
vec<-result$vectors
sigma1demi<-round(vec%*%val%*%t(vec),2)
epsilon<- rep(0,n)

# epsilon est un vecteur où on va stoquer les valeurs calculées

Z<-rnorm(n)
epsilon <-sigma1demi%*%Z
epsilon }

#################   PROGRAMME PRINCIPAL    ################### #
#on génère le couple(X,Z) de loi normale et de corrélation rho
#set.seed(3) U<-normal.bivariee(n,rho) X <- U[1,] Z <- U[2,]
compteur<-rep(0,5)

# On génère 2000 vecteurs Y de taille "n"
# pour garder la trace de la simulation des erreurs epsilon  on fixe le germe

set.seed(2009) for(k in 1:nsimul)
  {
  loi<-match.arg(loi)
  if (loi=="normaldep") epsilon<- multinormale.dep(n)
         else if (loi=="uniforme") epsilon<- runif(n,-1,1)
```

```
 Y<- beta1*X + beta2*Z + epsilon

# On estime le paramètre inconnu par la méthode du bootstrap

  coef<-matrix(0,nrow=1000,ncol=2)
  coef[1,]<-coefficients(lm(Y~Z))
  for (k in 2:1000){
      s<-sample(n,replace=T)
      Ys<- Y[s]
      Zs<- Z[s]
      coef[k,]<-coefficients(lm(Ys~Zs))
     }
  coef.btsp1<-mean(coef[,1])
  coef.btsp2<-mean(coef[,2])

  # la statistique du test est la corrélation partielle entre Ryx.z
  # on fait la régression linéaire du vecteur Y sur Z
  # et la régression linéaire du vecteur X sur Z

  Ryz<-residuals(lm(Y~Z))
  Rxz<-residuals(lm(X~Z))
  R_obs <- (sum(Ryz*Rxz)^2)/(sum(Ryz^2)*sum(Rxz^2))
  p.ttest<- coefficients(summary(lm(Y ~ X+Z)))[2,4]
  coef<-coefficients(lm(Y~Z))

  # la statistique du test  est déterminée par l'utilisateur
  # elle calcule la corrélation entre les deux vecteurs de  résidus
  # R_obs est la valeur  observée de la statistique du test
  # on définie des vecteurs pour stoker des calculs de la régression
```

```
stat.freedman <- numeric(nperm)
stat.kennedy <- numeric(nperm)
stat.manly <- numeric(nperm)
stat.propose <- numeric(nperm)
Res.freedman<-numeric(n)
Res.manly<-numeric(n)
Res.propose<-numeric(n)

# On estime le paramètre inconnu par la méthode du bootstrap
# cette boucle calcule la distribution empirique de la statistique
#  du test pour les différentes méthodes

for (i in (1:nperm)) {

     # on permute le vecteur des résidus Ryz  et le vecteur Y
     # on calcule la statistique de Kennedy directement
     # on régresse la variable de réponse Y permutée sur Z  (Manly)

     s<-sample(n)
     RYZpi<- Ryz[s]
     Ypi<- Y[s]
     stat.kennedy[i] <- (sum(RYZpi*Rxz)^2)/(sum(Rxz^2)*sum(Ryz^2))
     Res.manly<-residuals(lm(Ypi~Z))
     stat.manly[i]<- (sum(Res.manly*Rxz)^2)/(sum(Rxz^2)*sum(Res.manly^2))

     # on utilise les résidus permutés pour construire des nouvelles
     # observations de Ypi
     # on régresse les nouvelles observations Ypi sur Z et on calcule la
     # statistique de freedman et la méthode
```

```
        Ys<-coef.btsp1 +coef.btsp2*Z +  RYZpi
        Res.propose <- residuals(lm(Ys~Z))
        stat.propose[i]<-
        (sum(Res.propose*Rxz)^2)/(sum(Rxz^2)*sum(Res.propose^2))
        Ypi<- coef[1] + coef[2]*Z + RYZpi
        Res.freedamn<-residuals(lm(Ypi~Z))
        stat.freedman[i]<-
        (sum(Res.freedamn*Rxz)^2)/(sum(Rxz^2)*sum(Res.freedamn^2))
        }

    # on fait le tri des  distributions nulle calculées
    # calcule la puissance pour un test bilatéral

    null.stat.freedman <- sort(stat.freedman)
    null.stat.kennedy <- sort(stat.kennedy)
    null.stat.manly <- sort (stat.manly)
    null.stat.propose <- sort (stat.propose)
    pvalue.freedman<-(sum(null.stat.freedman>= R_obs)+1)/(nperm+1)
    pvalue.kennedy<-(sum(null.stat.kennedy>= R_obs)+1)/(nperm+1)
    pvalue.manly<-(sum(null.stat.manly >= R_obs)+1)/(nperm+1)
    pvalue.propose<-(sum(null.stat.propose >= R_obs)+1)/(nperm+1)
    # On compte le nombre de fois qu'on rejette l'hypothese nulle
    # on compare la pvalue pour chaque méthode avec le niveau alpha=0.05
    if (pvalue.freedman <= 0.05) compteur[1]<- compteur[1] + 1
    if (pvalue.kennedy <= 0.05)  compteur[2]<- compteur[2]+ 1
    if (pvalue.manly <= 0.05)  compteur[3]<- compteur[3] + 1
    if (pvalue.propose <= 0.05)  compteur[5]<- compteur[5] + 1
    if (p.ttest <= 0.05) compteur[4]<-compteur[4]+ 1
    }
puissance<-compteur/nsimul puissance }
```

```
#####################  TEST DE MANTEL PARTIEL   ####################

partial.mantel<-function(M1,M2,M3) { # les entrées de la fonction
partial.mantel sont des matrices de distances  qu'on suppose
symétrique avec diagonales nulles
   n<- dim(M1)[1]
   nperm<-4999
   Y<-as.vector(as.dist(M1))
   X<-as.vector(as.dist(M2))
   Z<-as.vector(as.dist(M3))
# on fait le test usuel d'un coefficient de régression. la
régression linéaire de la partie inférieure de la matrice Y  sur
les éléments de la partie inférieure des les matrices X et Z
   p.ttest<- coefficients(summary(lm(Y ~X+Z)))[2,4]
   res1<-residuals(lm(Y~Z))
   res2<-residuals(lm(X~Z))
   r_carrée<- (sum(res1*res2)^2)/(sum(res1^2)*sum(res2^2))
# on stocke les residus de Y  sur Z dans la matrice D
   coef<-coefficients(lm(Y~Z))
   D_yz<-matrix(0,n,n)
   D_yz[row(D_yz)>col(D_yz)]<-res1
   D<-D_yz +t(D_yz)
# on cree des vecteurs pour stocker la distribution de r^2 selon la
technique  de permutation utilisée
   stat.manly<-numeric(nperm)
   stat.kennedy<-numeric(nperm)
   stat.freedman<-numeric(nperm)
   for (i in 1:nperm){
       s<- sample(n)
       # on permute la matrice dépendante M1 et la matrice des résidus D
```

```
        D_pi<- D[s,s]
        # La méthode de Manly
        M1_pi<- as.vector(as.dist(M1[s,s]))
        res.manly<- residuals(lm(M1_pi~Z))
        stat.manly <- (sum(res.manly*res2)^2)/(sum(res.manly^2)*sum(res2^2))
        # la méthode de Kennedy
        res_pi<-as.vector(as.dist(D_pi))
        stat.kennedy<- (sum(res_pi*res2)^2)/(sum(res_pi^2)*sum(res2^2))
        # la méthode de Freedman
        Y.freedman <- coef[1] +coef[2]*Z+ res_pi
        res.freedman<- residuals(lm(Y.freedman~Z))
        stat.freedman<- (sum(res.freedman*res2)^2)/(sum(res.freedman^2)*sum(res2^2))


  }
   null.stat.freedman <- sort(stat.freedman)
   null.stat.kennedy <- sort(stat.kennedy)
   null.stat.manly <- sort (stat.manly)

   pvalue.freedman<-(sum(null.stat.freedman>= r_carrée)+1)/(nperm+1)
   pvalue.kennedy<-(sum(null.stat.kennedy>= r_carrée)+1)/(nperm+1)
   pvalue.manly<-(sum(null.stat.manly >= r_carrée)+1)/(nperm+1)

list(r_carrée,pvalue.manly,pvalue.freedman,pvalue.kennedy,p.ttest)
}
```

RÉFÉRENCES

Anderson, M.J. and Robinson, J. (2001). « Permutation tests for linear models ». *Aust. N. Z. J. Stat*, vol. 43(1), p.75-88.

Anderson, M.J. and Legendre, P. (1999). « An Empirical comparaison of permutation methods for tests of partial regression coefficients in a linear model ». *J. Statst. Comput. Simul*, vol. 43(1), p.75-88.

Anderson, M.J. (2001). « Permutation tests for univariate or multivariate analysis of variance and regression ». *Canadian Journal of Fisheries and Aquatic Sciences* , vol.58, p.626-639.

Cade, B.S. and Richards, J.D. (1994). « Permutation tests for least absolute deviation regression ». *Biometrics*, vol.52, p.886-902.

Cade, B.S. (2005). « Linear models : permutation methods ». *Encyclopedia Of Statistics In Behavioraal Science* , vol.2, p.1049-1054.

Dietz, P.E.J. (1983).« Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique »*Biometrics*, vol.34, p.277-282.

Doerge, R.W. and Churchill, G.A. (1996). « Permutation tests for multiple loci affecting a qunatitative character ». *Genetics*, vol.142, p.285-294.

Draper, D., Hodges, J.S. and Mallows, C.L.B.S. (2005).« Exchangeability and data analysis ». *Journal Of Royal Statistical Society* , vol.156, p.9-28.

Ernest, M.D. (2004). « Permutation methods : a basis for exact inference ». *Statistical Science*, vol.19,p.976-685.

Freedman, D. and Lane, D.(1983). « A Nonstochastic interpretation of reported signification levels ». *Journal Of Business And Economic Statistics*, vol.1(4), p.292-298.

Good, P. (1994). *Permutation tests : a practical guide to resampling methods for testing hypotheses*, New York. Springer Verlag

Good, P. (2002). « Extensions of the concept of exchangeability and their applications ». *Journal Of Modern Applied Statistical Methods* , vol.1, p.243-247.

Kennedy, P.E. (1995). « Randomization tests in econometrics ». *Journal of Business and Economic Statistics* , vol.25(4), p.923-936.

Kennedy, P.E. and Cade ,B.S. (1996). « Randomization tests for multiple regression ». *Commun. Statist-Simula*, vol.25(4), p.923-936.

Hogarty, K.Y. and Kromrey, J.D. (2003). « Permutation Tests For Linear Models In Meta-Analysis : Robustness And Power Under Non-Normality And Variance Heterogeneity ». *American Educational Research Association*.

Legendre, P. (2000). « Comparison of permutation methods for partial correlation and partial mantel tests ». *J. Statst. Comput. Simul*, vol.67, p.37-73.

Manly, B.F.J. (1997). *Randomization, bootstrap and monte carlo methods in biology*. Second edition. London : Chapman & Hall.

Mantel, N. (1967). « The detection of disease clustering and generalized regression approch ». *Cancer Research*, vol.27,part 1, p.209-220.

Mantel, N. and Valand, R.S. (1970). « A Technique of nonparametric multivariate analysis ». *Biometrics*, vol.26, p.547-558.

Mielke, P.W. (1978).« Permutation Tests For Associations Between Two Distance Matrices »*Syst. Zool*, vol.32(1), p.21-26.

Pesarin, F. (2001). *Multivariate permutation tests : with application in biostatistics*. Wiley, New York.

Randles, R.H. and Wolfe, D.A. (1979). *Introduction to the theory of nonparametric statistics* . J. Wiley, New York.

Randles, R.H. (1984). « On Tests applied to residuals ». *Journal of the American Statistical Association* , vol.79, p.349-354.

Shannon, W.D., Watson, M.A., Perry, A., and Rich, K. (2002). « Mantel statistics to correlate gene expression levels from microarrays with clinical covariates ». *Genetic Epidemiology*, vol.23, p.87-96.

Schmoyer, R.L. (1994). « Permutation tests for correlation in regression errors ». *Journal Of The American Statistical Association*, vol.89, p.428-437.

Smoose, P.E., Long, J.C. and Sokal, R.R. (1986). « Multiple regression and correlation extensions of the mantel test of matrix correspondence ». *Syst. Zool*, vol. 35(4), p. 627-632.

Wald, A. and Wolfowitz, J. (1944). « Statistical tests based on permutations of the observations ». *The Annals Of Mathematical Statistics*,Vol.15(4), p.358-372.

Welsh, W.J. (1990). « Construction of permutation tests ». *Journal of the American Statistical Association* , vol.85, p.693-698.

Welsh, W.J. (2003). « Transformations which preserve exchangeability and application to permutation tests ». *Journal of Nonparametric Statistics* , vol.15, p.171-185.