Partie : exploration des données

Pas besoin d’apprendre les formules mais c’est un petit +

**1. ACP:**

**Expliquer comment et pourquoi on passe d'un tableau de Big data que l'on souhaite explorer à la matrice de variance-covariance (ou matrice de dispersion), puis à celle de corrélation. Expliquer comment on en arrive à considérer les vecteurs propres de cette matrice de dispersion comme les directions des axes les plus propices à la conservation de la plus grande inertie et à une exploration pratique de données projetées de grandes dimensions.**

L’idée du problème est que nous nous retrouvons avec un tableau de big data immense avec des dizaines de colonnes. On sait faire des reg corr., des Anova, des tests du chi carré, mais on ne va pas passer notre temps à essayer toutes les combinaisons possibles… On va chercher un point de départ pour réussir à soutirer de l’informations sur bases des données qui nous sont proposées.

On va donc partir de 2 groupes de méthodes :

D’une part les méthodes factorielles et d’autre part les méthodes de classification.

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

On a un plan de base avec des données fournies et caractérisée par un nombre n de variables.

Dans le cas de l’ACP, on ne s’occupera que de variables quantitatives(taille, poids, etc.), pour chaque individu, sa représentation dans le schéma varie selon la variable, on représente cette individu dans un espace à n dimensions en fonction du nombre de variables quantitatives donc.

Mais nous ne pouvons voir qu’un espace à 3 dimensions donc on va devoir réduire notre nuage de points d’un espace à n dimensions sur un plan à 2 dimensions. (Exemple avec les bosses du chameau de profil, de face et de derrière).

On veut donc le meilleur plan possible pour visualiser nos données et malgré cela, il sera impossible de ne pas perdre de l’information en chemin évidemment. Pour faire parler nos données, pour en retirer de l’information, il faut accepter en perdre !

Une image contenant capture d’écran, ligne

Description générée automatiquementAprès la projection, on va se rendre compte que certains points vont se retrouver ensemble, alors que d’autres seront, au contraire, très éloignés. Si notre nuage de points représente les individus, on va pouvoir regarder les groupes d’individus et s’intéresser à pourquoi ils se retrouvent ensemble.

Deux points proches l’un de l’autre, on peut en déduire que ces deux variables sont corrélées. Et on verra plus tard les moyens de vérifier que ces deux variables sont effectivement bien corrélées.

Remarque lorsque l’on regarde notre projection, au grand jamais il ne faut regarder l’information centrale qui est biaisée.

Pour en revenir aux méthodes factorielles et de classification :

La première servira à trouver des formules, de régression par exemple, tandis que la seconde est là pour chercher des groupes avec la vue la plus simple sans modèles ou formules.

On veut donc trouver les 2 meilleurs axes pour notre ACP(à 2 dimensions) et pour cela on cherche la projection et qui déforme le moins possible le nuage initial.

Une image contenant ligne, diagramme, Tracé, capture d’écran

Description générée automatiquementOn peut même pour cela mettre en 1 dimension notre plan (réduction de l’information mais plus grande précision dans la dispersion des points).

On doit aussi calculer ses axes pour trouver les meilleurs, donc en partant de notre plan à n dimensions et cela nous conduire au meilleur plan tout en minimisant la perte d’inertie.

Tableau de données -> matrice

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquementSi on considère notre table avec p variables quantitatives : X1,X2,X3… et n individus. On se retrouve avec n\*p observations Xij et Bam notre matrice est faite.

Une image contenant diagramme, capture d’écran, ligne, cercle

Description générée automatiquementMaintenant pour chercher les axes parfaits, on veut calculer l’inertie et pour cela on regarde la distance entre les différents points.

On peut aussi calculer le point de gravité du plan qui en fait la moyenne des distances entre chaque points d’individus.

Matrice -> matrice transformée (variables centrées)

On prend pour chaque point non plus sa valeur mais l’écart par rapport à leur moyenne respective (du centre de gravité). On peut faire le lien avec la loi normale ou la moyenne des points étaient de 0 et bah là c’est pareil pour notre centre de gravité qui devient l’origine (0,0).

On passe enfin à des valeurs centrées réduites pour s’affranchir de la dépendance du système unitaire. -> matrice des valeurs centrées réduites

Sur base de cette nouvelle matrice, on va pouvoir construire une matrice carrée afin de trouver l’inertie parfaite. Et donc il suffit de multiplier notre matrice par la même matrice transposée, donc le produit scalaire.

Avec cette matrice carrée, on peut obtenir :

* la matrice de covariance(dispersion) V = on multiplie la matrice centrée par sa transposée, on rajoute à la matrice le \*(1/n)
* la matrice de corrélation R = on multiplie la matrice centrée réduite par sa transposée, on rajoute à la matrice le \*(1/n)

Avec la matrice de covariance, on multiplie notre matrice centrée par sa transposée et par 1/n et on obtient une matrice carrée.

Si on prend les variances sur la diagonale de notre matrice de dispersion, on observe l’étalement du nuage sur toutes ces dimensions et donc l’inertie.

Sur R, pour trouver nos vecteurs parfaits, on va lui donner notre matrice de dispersion avec la fonction eigen(matrice) qui nous donnera les valeurs et vecteurs propres.

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

**2. ACP:**

**Expliquer comment on interprète la représentation des individus et l'origine de cette manière de faire. Comment peut-on alors déterminer la signification concrète des axes principaux F1 et F2 ? Qu'apportent dans ce contexte les résultats fournis par la fonction PCA de FactoMineR ?**

Explication de la représentation des individus :

Ce graphique résulte d'un nuage de points dans un espace à n dimensions, projeté sur un espace à 2 dimensions, avec les deux axes correspondant aux deux vecteurs propres les plus significatifs extraits de la matrice de dispersion. L'objectif est de regrouper des individus similaires et de déterminer la possibilité de former des groupes. Il est essentiel de noter que le graphique peut parfois être mal projeté, comme dans l'analogie d'un astronome observant deux étoiles dans une lunette, où elles peuvent sembler proches alors qu'elles sont en réalité éloignées en profondeur.

En disposant les individus le long de l'axe horizontal, ceux à gauche peuvent représenter des observations présentant des caractéristiques spécifiques, tandis que ceux à droite peuvent représenter d'autres caractéristiques. Le premier axe (F1) sépare ainsi les individus en fonction de certaines propriétés, expliquant un pourcentage donné de l'inertie totale, indiqué sous l'axe horizontal.

En examinant le deuxième axe (F2), les individus au-dessus peuvent avoir des caractéristiques relativement constantes, tandis que ceux en dessous peuvent montrer une variation significative dans une autre dimension.

Avec R, nous utilisons la fonction PCA() pour effectuer une Analyse en Composantes Principales. Nous examinons les valeurs propres pour chaque direction, choisissons le nombre optimal de vecteurs pour maximiser l'inertie, et visualisons les pourcentages d'inertie des vecteurs à l'aide de la méthode eig() et d'un barplot().

**3. ACP:**

**Expliquer comment on interprète la représentation des variables et l'origine de cette manière de faire. Que sont en fait les coordonnées des points qui représentent chaque variable ?**

**Quel rapport entre la qualité de la projection et le cosinus ?**

**Qu'apportent dans ce contexte les résultats fournis par la fonction PCA de FactoMineR ?**

**Expliquer pourquoi et comment on peut voir que des variables sont corrélées d'une manière ou d'une autre ou, au contraire, ne le sont pas du tout.**

Interprétation de la représentation des variables :

La représentation des variables dans l'ACP permet de visualiser la contribution de chaque variable à la variance des données. Les coordonnées des points représentant chaque variable sont importantes pour comprendre leur impact sur les composantes principales. Ces coordonnées indiquent la direction et l'amplitude de la variation de chaque variable dans l'espace des composantes principales.

Origine de cette manière de faire :

L'origine de cette approche réside dans la réduction de la dimensionnalité des données tout en préservant au mieux la variance totale. Elle permet de simplifier la compréhension des structures sous-jacentes des données en identifiant les principales sources de variation.

Coordonnées des points représentant chaque variable :

Les coordonnées des points représentant chaque variable indiquent leur contribution respective aux composantes principales. Plus la coordonnée d'une variable est éloignée de zéro, plus son impact sur la composante correspondante est fort.

Le cosinus de l'angle entre la variable originale et sa projection sur le plan des composantes principales mesure la qualité de la projection. Un cosinus proche de 1 indique une bonne projection, signifiant que la variable est bien représentée dans l'espace réduit. Un cosinus proche de 0 suggère une mauvaise projection.

- Angle petit => cosinus très grand positivement. Effectivement : cos(0) = 1 [Corrélation positive]

- Angle très grand => cosinus très grand négativement. Effectivement : cos(180) = -1 [Corrélation négative]

- Angle droit => cosinus nul. Effectivement : cos(90) = 0 [quadrature]

La corrélation entre variables peut être observée en examinant la proximité de leurs représentations dans l'espace des composantes principales. Si deux variables sont situées près l'une de l'autre, elles sont corrélées positivement ou négativement en fonction de leur orientation et du cosinus formé par l’angle des vecteurs(représentant 2 variables). À l'inverse, si elles sont éloignées, cela suggère une faible corrélation(schéma circulaire).

Avec R, on lit la table, on fait l’ACP via la méthode PCA(), avec ça on a accès aux coordonnées des variables, aux cosinus de qualité de projection, et aux contributions des variables.

On regarde nos valeurs propres, correspond à une direction. On prend le n de vecteurs pouvant nous apporter un maximum d’inertie et donc choisir le meilleur plan de projection, on visualise les pourcentages d’inerties des vecteurs via la méthode $eig() mais aussi en schéma via un barplot de notre ACP.

La projection à proprement parlée est le produit scalaire entre 2 vecteurs, on prends la longueur du premier que l’on multiplie par la longueur du second, le tout multiplier par le cosinus de l’angle entre les deux.

**4. ACM:**

**Expliquer les principes de construction des nuages de point. En particulier, comment parvient-on au nuage de points des modalités et comment tient-on compte du fait que certaines modalités sont rares ?**

L’ACM est aux données qualitatives ce que l’ACP est aux données quantitatives. + On n’oublie pas que les valeurs, dans le cas des variables qualitatives sont appelées des ‘modalités’.

Nous prendrons comme exemple ici un tableau représentant une étude de satisfactions de 28 clients dans un garage automobile. Les 5 variables sont qualitatives (! Accueil n’est pas du quanti, c’est un niveau de satisfaction – bien reçu, très bien reçu – déplorable, … L’échelle entre les valeurs n’est pas respectée, c’est donc bien du qualitatif).

Une image contenant texte, nombre, reçu

Description générée automatiquement**rep : NSP, OUI, NON et Accueil : 0 - 5**

On cherche à trouver des associations entre variables, Mais plus précisément, on ne cherche pas des relations entre les variables en elles-mêmes mais des relations entre les modalités, est ce que ‘modalité1’ a une influence sur ‘modalité2’ ?

Le but va être de détecter les individus ayant le même profil : on cherche donc à mesurer la distance entre les individus. 2 personnes ayant les mêmes modalités auront une distance nulle. 2 personnes ayant des modalités semblables auront une distance faible.

Mais si sur ces 2 individus, une seule modalité change, et que celle-ci est une modalité rare, alors là, la distance sera grande. On cherche les modalités rares.

Et a l’inverse, si deux clients ont des modalités fort différentes, mais qu’ils ont en communs la même modalité rare, alors eux seront proches.

**Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement**

L’ACM, elle, va nous donner 3 dessins. Comme pour l’ACP, un pour les individus et un pour les variables.

Mais en plus, un troisième sans équivalent, le diagramme des modalités ! Et c’est probablement le plus intéressant.

De ces modalités, on veut en faire ressortir **un tableau disjonctif complet(TDC)**, donc on va encore réaliser une matrice n (individus)\*p (variables), Dans ce tableau, chaque variable qualitative j peut prendre plusieurs modalités k et p´ colonnes donnant : p´ = SUM(Kj).

Une image contenant Police, texte, nombre, ligne

Description générée automatiquement

Avec ce tableau, on va pouvoir calculer des proportions, fréquences, probabilités pour une modalité donnée. On peut calculer le % d’un choix d’une modalité possible en prenant le n de réponse d’une modalité et en la divisant par le nombre d’individus intervenants.

On calcul ces probabilités car, les 2 objectifs de l’ACM sont :

* Montrer ce qui va ensemble : donc il faut mesurer une distance.
* Montrer des modalités qui sortent de l’ordinaire (car l’ordinaire ne m’apprend rien).

Et les modalités rares, ont une faible probabilité donc il faut donc qu’on les calcule.

**5. ACM:**

**Décrire la marche à suivre en R et les règles à appliquer pour interpréter les graphiques résultant et les informations fournies par la fonction MCA (pourquoi les valeurs propres, les rapports de corrélation(cos²), ?).**

Tout d’abord on va compiler la librairie FactoMineR. On lit les données du fichier en question et c’est parti. On transforme le type des colonnes voulues en qualitatives avec as.factor() si ce n’est pas le cas.

Puis on utilise la méthode MCA() pour obtenir notre ACM, en faisant acm$eig cela nous donne comme pour l’ACP, les valeurs propres, le pourcentage d’inertie selon chaque axe et l’inertie cumulée.

Une image contenant texte, capture d’écran, Police, reçu

Description générée automatiquement

Avec ce tableau, on va pouvoir décider de combien de vecteurs propres on veut garder pour expliquer l’inertie en regardant le pourcentage de ces vecteurs cumulées. Une fois choisi on peut effectuer la méthode suivante en précisant les vecteurs(dimensions) que l’on veut.

Acm$cos2() => Pour savoir si une variable est bien projetée ou non, on va regarder les cos², ils représentent la qualité de la représentation des points et des variables sur les dimensions. Les cos² proches de 1 indiquent une bonne représentation, tandis que ceux proches de 0 indiquent une mauvaise représentation.

Cosinus grand = quand l’angle formé entre la variable et l’axe est consideré petit, donc qu’ils sont proches.

**Une image contenant texte, Police, document, capture d’écran

Description générée automatiquement**

Et les méthodes pour afficher les représentation des variables et individus :

Une image contenant texte, Police, calligraphie, typographie

Description générée automatiquement

**6. ACP/ACM:**

**Comparer globalement les deux techniques d'exploration des données ACP et ACM basées sur les projections de nuages de points : contenu de ces nuages de points, principe directeur de la projection, graphiques résultants et données numériques disponibles (pourquoi les valeurs propres, les cos2, les coefficients/rapports de corrélation) après utilisation de la fonction R (laquelle ?) correspondante.**

1. **Contenu des nuages de points :**
   * **ACP (Analyse en Composantes Principales) :**
     + Les nuages de points en ACP représentent la projection des individus et/ou des variables dans un nouvel espace, appelé espace des composantes principales.
     + Les individus sont projetés sur les axes principaux qui maximisent la variance des données.
     + Les variables sont projetées sur les mêmes axes, indiquant leur contribution à la variance totale des données.
   * **ACM (Analyse en Correspondances Multiples) :**
     + Les nuages de points en ACM représentent la projection des lignes et des colonnes d'un tableau de contingence dans un nouvel espace.
     + Les lignes (modalités des variables qualitatives) et les colonnes (modalités des individus) sont projetées sur des plans factoriels.
2. **Principe directeur de la projection :**
   * **ACP :**
     + Le principe directeur est de maximiser la variance des données dans les premières composantes principales.
     + Les premières composantes principales expliquent la majeure partie de la variance, permettant ainsi une réduction de dimension.
   * **ACM :**
     + Le principe directeur est d'analyser les associations entre les modalités des variables qualitatives.
     + Les plans factoriels révèlent des associations entre les lignes et les colonnes du tableau de contingence.
3. **Graphiques résultants :**
   * **ACP :**
     + Les graphiques incluent des nuages de points des individus, des cercles de corrélation des variables, etc.
     + Ils permettent de visualiser la structure des données dans l'espace des composantes principales.
   * **ACM :**
     + Les graphiques incluent des nuages de points des lignes et des colonnes dans les plans factoriels.
     + Ils mettent en évidence les associations entre les modalités des variables qualitatives.
4. **Données après utilisation de R :**
   * **ACP (avec R, par exemple en utilisant la fonction MCA du package FactoMineR) :**
     + Les valeurs propres : indiquent la variance expliquée par chaque composante principale.
     + Les cos2 (carrés des cosinus) : mesurent la qualité de la représentation des individus et des variables dans l'espace des composantes principales.
   * **ACM (avec R, par exemple en utilisant la fonction MCA du package FactoMineR) :**
     + Les valeurs propres : indiquent la variance expliquée par chaque axe factoriel.
     + Les cos2 : mesurent la qualité de la représentation des lignes et des colonnes dans les plans factoriels.
     + Les coefficients/rapports de corrélation : indiquent la corrélation entre les modalités des variables qualitatives et les axes factoriels.

**7. HCPC:**

**Expliquer les principes de base de construction d'une classification/partition d'un groupe d'individus selon une CAH. Décrire les principaux types de distances entre individus et de méthodes d'agrégation entre classes que l'on peut utiliser pour créer une partition selon le principe de la classification hiérarchique ascendante. En particulier, comparer l'effet de l'utilisation de la méthode d'agrégation de Ward avec d'autres méthodes plus simples.**

**Une image contenant texte, capture d’écran, Police, logiciel

Description générée automatiquement**

On parle d’un système de classification, on a plusieurs groupes de + en + grands si on part de tout en bas et l’inverse si on part du groupe le plus haut(classification ascendante hiérarchique).

**Une image contenant texte, capture d’écran, diagramme, conception

Description générée automatiquement**

**Une image contenant texte, capture d’écran, nombre, affichage

Description générée automatiquement**

On calcule la distance entre les points en cherchant ceux avec la distance la plus courte, on fait des agrégations jusqu’à n’avoir plus qu’un groupe.

Notre algorithme dépends de différents paramètres :

* La manière de calculer les distances (euclidienne, city-block)
* Le choix de comment les rassembler (saut min, saut max, moyenne, Ward)

Donc on connait plusieurs moyens pour regrouper, la méthode du saut minimal est de regrouper chaque points ayant la distance la plus petite entre eux jusqu’à ce qu’il ne reste qu’une seule classe(groupe), utile lorsqu’on l’utilise sur des groupes très distincts(Ward aura les mêmes résultats que celle-ci).

Dans ces moyens on a vu aussi la distance de Ward qui est plus efficace que celle du saut minimal. Elle permet de faire des agrégations entre groupes en limitant la perte d’inertie intra-classe et minimiser l’augmentation de l’inertie inter-classe. Elle est intéressante lorsque l’on doit travailler avec des groupes qui se confondent étroitement.

Une image contenant texte, capture d’écran, Police, Parallèle

Description générée automatiquement

**8. CAH:**

**Décrire les graphiques fournis par une CAH classique - pourquoi parle-t-on de *coupure* ?**

**Expliquer les coefficients q et AC - en quoi permettent-ils d'apprécier une partition en plus ou moins discriminante ?**

**HCPC: Expliquer les principes d'une CAH appliquée à une ACP ou une ACM.**

**Quelles sont les ressemblances et les différences avec une CAH classique ? Quelles informations intéressantes peut-on trouver dans les clauses "$data.clust", "$desc.var" et "$desc.ind" du résultat fourni par la fonction HCPC() de R ?**

**Expliquer le principe de la méthode de classification/partitionnement des Kmeans. Quel intérêt a-t-on à appliquer la méthode des K-means sur les résultats d'une HCPC basée sur une ACP ou une ACM, ou encore directement sur les résultats d'une ACP ou ACM ?**

**CAH**

Une image contenant texte, capture d’écran, Parallèle, nombre

Description générée automatiquement

**= La matrice des dissimilarités.**

Le coefficient d’agglomération (AC) est là, comme q tout à l’heure, pour mesurer le fait qu’une partition est bonne, ou mauvaise. Plus notre AC sera grand, plus la classification sera discriminante (groupes bien séparés, avec des individus proche intra-groupe).

* AC proche de 1 : OK.
* AC proche de 0 : pas bon.

On a ce qu’on appelle la classification ascendante hiérarchique. On va pouvoir rassembler des groupes selon leur points communs. Nous ce que l’on veut là c’est quand on a déjà les groupes formés et que l’on veut petit à petit les séparer en petits groupes, la découpe en arbre.

La question de savoir combien de groupes on veut garder et une fois cela fait on coupe à l’endroit décidé.

**Une image contenant texte, Police, ligne, capture d’écran

Description générée automatiquement**

Une image contenant diagramme, ligne, Rectangle, Plan

Description générée automatiquement**HCPC**

Une image contenant texte, Police, capture d’écran, typographie

Description générée automatiquement

On appelle donc la fonction HCPC() à qui on file les résultats et qui nous construit un dendrogramme basé sur les données projetées amenées par l’ACP. Mais ce n’est pas tout… HCPC() attend de nous qu’on clique sur le dessin pour ‘couper’ notre arbre.

Pour ce faire, il nous donne le plot de dénivellations du nombre de classe. Ainsi, on sait si notre découpage conserve des classes fort discriminantes ou non.

Une image contenant ligne, Rectangle, capture d’écran, diagramme

Description générée automatiquementCe graphique peut être interprété tel que, plus on avance :

- Plus on perd de l’inertie inter-classe. (Évidemment, au départ elle était maximale), plus on fait de regroupement et plus cette inertie diminue jusqu’à devenir nul au dernier groupement.

- Plus on gagne de l’inertie intra-classe

Petite remarque ici, l’arbre obtenu ici est fort aplatit en bas, logique. Il ne faut pas oublier qu’on travaille sur une projection de p dimensions sur un plan à simplement 2 dimensions.

Bref, on clique sur le dessins pour couper notre arbre. Et dès qu’on coupe, il sort ce dessin-là :

Une image contenant texte, diagramme, capture d’écran, ligne

Description générée automatiquementIl nous montre directement le diagramme des individus coloriés en fonctions des différents groupes auxquelles ils appartiennent.

On obtient également un autre dessin à dimensions qui permet de voir les groupement ainsi que les hauteurs, mais pas franchement lisible.

On récupère lors d’une HCPC, une variable contenant plusieurs informations.

$data.clust : qui permet de nous indiquer à quel cluster (à quel groupe) appartiennent chacun de nos individus.

$dest.var : qui nous sort toute une série d’Anova.

$desc.ind:

* $desc.ind$para : Nous donne le parangon de chaque classe. (L’individu représentant au mieux le groupe auxquels il appartient. C’est l’individu le plus proche du centre de gravité de la classe.
* $desc.ind$dist : Fait l’inverse, trie la liste de chaque classe en commençant par les plus éloignés et en tête de liste l’élément le plus distinctif de sa classe.

Pour ce qui en est du **qualitatif**, on lui fait passer dessus une ACM qui nous donne un nuage projeté d’individus, donc, comme pour l’ACP, on peut lui faire passer dessus un HCPC et on peut en soutirer nos informations de groupes basée sur des variables qualitatives.

Pour faire ça sur à la fois des variables quantitatives et qualitatives, on utiliserait la fonction FDMA().

**Kmeans**

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

**9. CAH:**

**Quels sont les différents critères d’agrégation pour une CAH ? Expliquez à l’aide d’un exemple.**

Utile dans une hiérarchie avec des groupes d’individus. C’est le fait de rassembler plusieurs individus de groupes différents en un seul.

Il y a différents critères d’agrégation :

En Classification Ascendante Hiérarchique (CAH), les critères d'agrégation sont des méthodes qui déterminent comment les similarités ou distances entre groupes d'individus ou de variables sont agrégées pour former des groupes plus grands. Il existe plusieurs critères d'agrégation, chacun avec ses propres caractéristiques. Les critères d'agrégation couramment utilisés comprennent :

1. Critère de la distance minimale (Single Linkage) :
   * **Principe :** Lorsque deux groupes sont à fusionner, la distance entre ces groupes est définie comme la plus petite distance entre deux individus, l'un de chaque groupe.
   * Une image contenant cercle, diagramme

     Description générée automatiquement**Exemple :** Supposons que les distances entre les individus A, B, C et D soient respectivement 1, 3, 2, 5. La distance minimale entre les groupes {A, B} et {C, D} serait de 2, car C (du premier groupe) et D (du deuxième groupe) ont la distance minimale.
2. Critère de la distance maximale (Complete Linkage) :
   * **Principe :** La distance entre deux groupes est définie comme la plus grande distance entre deux individus, l'un de chaque groupe.
   * **Exemple :** En reprenant les distances de l'exemple précédent, la distance maximale entre les groupes {A, B} et {C, D} serait de 5, car D (du deuxième groupe) et A (du premier groupe) ont la distance maximale.
3. Critère de la distance moyenne (Average Linkage) :
   * **Principe :** La distance entre deux groupes est définie comme la moyenne des distances entre tous les individus des deux groupes.
   * **Exemple :** Si les distances entre A, B, C et D sont respectivement 1, 3, 2, 5, la distance moyenne entre les groupes {A, B} et {C, D} serait (1 + 3 + 2 + 5) / 4 = 11 / 4 = 2,75.
4. Critère de Ward :
   * **Principe :** Minimise la somme des carrés des écarts à la moyenne à l'intérieur de chaque groupe lors de l'agrégation. Ward vise à minimiser la variance globale après la fusion de deux groupes. Si la variance intra-groupe est plus faible après la fusion, alors c'est un bon candidat pour l'agrégation.
   * **Exemple :**

Valeurs initiales :

* + A: 3 -- B: 7 -- C: 11 => La moyenne initiale est (3+7+11)/3 = 7.

Maintenant, le critère de Ward voudrait minimiser la somme des carrés des écarts à la moyenne. Calculons ces écarts carrés pour chaque point :

Écarts carrés :

* + A : (3-7)^2 = 16
  + B : (7-7)^2 = 0
  + C : (11-7)^2 = 16

La somme des carrés des écarts est donc 16 + 0 + 16 = 32.

Maintenant, supposons que nous agrégeons les points A et B en un nouveau groupe AB avec une moyenne pondérée (3*1 + 7*1) / (1+1) = 5.

Nouvelles valeurs :

* + AB : 5
  + C : 11

Calculons à nouveau les écarts carrés :

Nouveaux écarts carrés :

* + AB : (3-5)^2 = 4
  + C : (11-5)^2 = 36

La somme des carrés des écarts après l'agrégation est donc 4 + 36 = 40.

Comparons les deux résultats :

* + Avant l'agrégation : 32
  + Après l'agrégation : 40

Dans ce cas, le critère de Ward déciderait de ne pas agréger A et B, car l'agrégation a augmenté la somme des carrés des écarts à la moyenne.

**10. REG-COR MULTIPLE:**

**Qu'entend-on par "régression-corrélation multiple" (exemple, mise en forme théorique de la régression et de la corrélation). Expliquer le rôle de l'utilisation de statistiques de Student et d'une statistique de Fisher dans l'analyse inférentielle d'une étude de régression-corrélation multiple.**

Lors d’un problème de régression corrélation, on s’intéresse à une variable quantitative en fonction d’une autre variable quantitative. Par exemple, la dextérité manuelle qui dépend de la vision lm(dexterite~vision).

On se rappelle la formulation mathématique (reg corr simple) : yi = β ^ 0 + β ^ 1 xi (+ εi)

Qui devient (reg corr multiple) : yi = β ^ 0 + β ^ 1 x1i + β ^ 2 x2i + β ^ 3 x3i + … (+ εi)

Le lm() fait de la statistique descriptive, il calcul a et b (moindre carré) avec les données de l’échantillon.

Le summary() fait de l’inférence statistique. Peut-on raisonnablement faire confiance aux valeurs estimées pour nos paramètres de notre modèle β ^ 0, β ^ 1, β ^ 2, β ^ x . Le summary() fait des tests de student. => H0 : β ^ x = 0.

Nous observons des p-values très petites, on rejette H0. Nos paramètres sont significativement différents de 0. On peut donc faire confiance aux valeurs estimées.

****

Le summary() nous donne également un statistique de Fisher ainsi que sa p-value associée, ce qui nous permet de savoir si oui ou non notre modèle nous apporte quelque chose. Pour cela :

On va prendre pour chaque xi, les y^ i calculé par notre modèle complet, cela nous donne un groupe de valeurs. Ensuite pour chaque xi, les yi observés (modèle vide), cela nous donne un 2ème groupe de valeurs. On réalise ensuite une Anova (donc Fisher) entre les valeurs prédites et les valeurs qui ne tiennent absolument pas compte du modèle et on regarde s’il y a une différence significative. Si oui : le modèle amène bien quelque chose.

Les gens qui ont fabriqué la reg corr multiple ont démarré avec des équations telles que :

SCE totale = SCE résiduelle + SCE factorielle



On sent très bien une combine du même genre que l’Anova. On repart avec des écarts factoriels et résiduels, des sommes, des carrés, etc…

Voilà ce qui explique entre autres pourquoi R nous donne un Fisher avec une reg corr.

On se rappelle aussi du coefficient de corrélation r [-1, +1] qui nous informait de s’il s’agissait d’une bonne idée ou non de se servir de la droite, de la courbe que l’on avait calculé. Est-ce que notre droite colle assez bien à notre nuage de point, ou alors est-ce bof… Le coefficient de corrélation r possède une formule générale, mais également une formule particulière pour le cas de la régression linéaire (COV(X, Y) / σx \* σy

Passons à la régression corrélation multiple. 

D’après notre modèle, notre formule, nous passons maintenant dans une espace à 3 dimensions, on ne parle plus de droite de régression mais de plan de régression (pour 3 variables observées).

Quand on s’attaque à une reg corr multiple, on va se retrouver avec plusieurs coefficient de corrélation.

* **Coeff corr d’ordre zéro :** Nous avons 3 variables, nous allons en prendre 2 parmi les 3 en oubliant la dernière et nous pouvons calculer le coeff corr de ces deux variables. On peut calculer un coeff corr d’ordre zéro pour chaque paire de variables : r12, r13, r23 . Cela se peut avérer utile lors de l’exploration de données.
* **Coeff corr partielle :** De nouveau on prend 2 variables parmi les 3 mais cette fois-ci, on n’oublie pas la 3ème, on la bloque sur une valeur constante. r12.3, r13.2, r23.1. La troisième variable intervient dans les calculs.

Dans ces 2 premiers cas, les 3 variables ont le même rôle, elles sont sur le même pied.

* **Coeff corr multiple :** (celui que nous donne r) Dans ce cas, les 3 variables n’ont PAS le même rôle. 1 joue le rôle de variable expliquée, et 2 jouent le rôle de variables explicatives. L’idée est que nous partons d’un dataframe avec deux colonnes, la première les yi observés, la deuxième les y^ i calculés par notre modèle, et c’est là-dessus que nous allons calculer notre coefficient de corrélation.

**Une image contenant Police, texte, écriture manuscrite, ligne

Description générée automatiquement**

Revenons en maintenant à notre paramètre de régressions :

Lors de l’analyse de l’échantillon, on peut trouver (par la méthode des moindre carrés) nos coeff a, b, c, d, … Mais si je change d’échantillon, évidemment ces valeurs vont varier. On se demande donc dans quel intervalle de confiance je peux prédire les valeurs de ces coefficients : On se demande donc comment se distribue nos paramètres grand A, grand B, grand C, grand D, … Une fois que l’on connait cette distribution, on prend la fourchette à 95%, c’est plié.

Le A et le B se distribuent selon une loi normale, mais R travaille toujours avec student (en effectuant un changement de variable). On passe donc par un t qui qui suit une loi de student.

Rappel lorsqu’on étudiait la régression corrélation simple.

Une image contenant texte, Police, ligne, nombre

Description générée automatiquement

Pour la multiple : t suit une loi de student à **n-m-1 DF**.

**Une image contenant Police, Graphique, logo, graphisme

Description générée automatiquement**On se rappelle que la loi de student est basée sur une centrée réduite. ((X – μ) / σ). On fait intervenir un écart-type dans formule. L’écart-type est construit à partir des écarts à la moyenne. Il y a une condition, la somme de ces écarts doit valoir 0. S’il y a n valeurs, il y a n écarts. Les n-1 premiers peuvent valoir ce qu’ils veulent mais le dernier doit s’arranger pour que la somme fasse zéro. Dans la reg corr simple il y a les variables X et Y (dataframe), on a toujours n valeurs mais avec 2 contraintes, on avait n-2 DF. Ici on a 3 variables X, Y et Z, on se retrouve avec n-3 DF.

Plus généralement, **n-m-1 DF** avec n l’effectif de l’échantillon, m le nombre de variables explicatives. 

**11. REG-COR MULTIPLE:**

**Décrire les étapes de traitement d'un problème de régression-corrélation multiple en R. En particulier, expliquer comment on peut réduire un modèle de corrélation multiple complet à un sous-modèle plus adéquat et plus fiable. Expliquer dans ce contexte ce qu'est un "coefficient de détermination ajusté".**

****

Nous voilà avec, on repart sur un modèle de genre là, avant autant de variables explicatives qu’on le souhaite (m).

Premièrement, on va devoir représenter nos données sous forme d’un tableau de big data, une colonne par caractère observé. On peut ensuite lire notre tableau avec un read.table().

**Une image contenant texte, Police, capture d’écran, nombre

Description générée automatiquement**On peut ensuite établir le modèle linéaire, qui va notre donner les estimations de nos paramètre b^ 0, b^ 1, b^ 2, b^ m calculés sur base de l’échantillon.

On passe maintenant au niveau de la population : **summary(lm())**

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquementOn retrouve comme à l’époque nos St pour vérifier si nos estimations de paramètres sont significativement égales ou non à zéro.

On se rend compte dans cet exemple, que **β0** et **β2** sont égales à nulle, ils ne servent à rien dans l’équation. On les laisse tomber. (Cela arrivera très souvent de laisser tomber notre ordonnée à l’origine). Mais on peut en plus dire ici que l’âge n’intervient pas dans notre équation de mesure du tour de poitrine. On commence à renifler un modèle ou le tour de poitrine serait lié au tour de taille mais où l’âge n’aurait rien à faire là-dedans.

Cela nous arrivera à plusieurs reprises, au départ on va tester un modèle global avec pas mal de régresseurs, pour se rendre compte que certains n’en valent pas la peine (significativement = 0), on passera donc à un sous-modèle avec moins de régresseurs qui interviennent.

Encore une fois, on peut observer le test de Fisher :



Et notre coefficient de détermination (notre R²) : et on détermine que notre modèle est honnête, on peut s’y fier.

Une image contenant texte, Police, calligraphie, Graphique

Description générée automatiquement

Logique, lorsque notre nombre de régresseurs augmentent, le R² va augmenter aussi. Imaginons des barres en métal de 5, 10 et 15 cm avec comme variation respective 0.2, 0.5, 0.4mm. Si on empile ses barres, on additionne ces tolérances, ces variations de mesures. Ici c’est pareil, à chaque variable qu’on amène, on amène des variations de plus en plus grandes… Et ces variations interviennent dans le calcul du coefficient de corrélation, il en va de soi que celui va également augmenter ! Imaginons un modèle avec énormément de régresseurs, on va se retrouver avec un R² magnifique, mais pour autant, la corrélation peut s’en avérée être totalement pourrie.

Une image contenant texte, ligne, capture d’écran, Police

Description générée automatiquementOn va donc calculer un coefficient de corrélation ajusté : on va le corriger de telle manière que, plus le nombre de régresseurs augmente, plus on diminue notre R². On observe un ‘-m’ dans la formule, m étant le nombre de variables. Plus il y en aura, plus grand sera le diviseur de la formule et ainsi plus petit sera le R² ajusté.

C’est donc ce R² ajusté qui nous informe du coefficient de corrélation. Attention, quand on regarde la formule, le R² ajusté, n’est pas un véritable carré, il garde le nom pour savoir de quoi on parle, mais si le nombre de régresseurs est suffisamment grand, on pourrait très bien se retrouver avec un R² ajusté négatif.



Enfin, comme dit précédemment, cela arrivera régulièrement de rejeter l’estimation de l’intercept lors du summary(), on ne tiens donc pas compte de la pente. Mais ATTENTION, elle est intervenue dans le calcul. Si elle est intervenue dans le calcul (lors du lm()), cela veut dire que nos estimations de β1 et β2 ne sont pas tout à fait juste.

Le ‘-1’ présent dans la formule lui précise qu’on ne souhaite pas s’occuper de la pente. En effet, si l’intercept est indéterminé (vu dans le summary() du modèle général), il va introduire une incertitude sur le calcul de la pente, on retire donc cette pente de notre formule et on recommence sans prêter attention ni à la pente, ni à l’intercept.

**12. ANOVA1/2:**

**Expliquer et comparer les tests d'hypothèses de l'ANOVA 1 et ANOVA 2 : position du problème et types de variables impliquées, bases mathématiques, hypothèse nulle et statistiques du test.**

L'ANOVA (Analyse de la Variance) est une méthode statistique utilisée pour étudier une variable quantitative en fonction d'une variable qualitative. On veut vérifier s’il y a une différence significative entre les moyennes de plusieurs groupes.

Lors de l'ANOVA 1, réalisée en R avec la fonction lm(), les résultats incluent la moyenne générale (µ) et les écarts des autres moyennes par rapport à la première (αi). Le summary(lm()) génère des tests de conformité, initialement H0 : µ = 0 pour l'intercept, suivi de H0 : αi = 0.

Le test H0 : µ1 = µ2 = µ3 dans l'ANOVA simple s'appelle l'analyse de la variance, car elle examine les écarts des valeurs par rapport à leurs moyennes, permettant la construction d'écarts-types et l'estimation de variances.

* H0 : Les moyennes des groupes sont égales (absence de différence significative).
* H1 : Au moins une moyenne de groupe est significativement différente.

La statistique de test Fisher est calculée comme le rapport de la variance intergroupes sur la variance intragroupes. Plus le ratio est élevé, plus il y a de chances que les moyennes des groupes diffèrent significativement.

Pour l'ANOVA2, l'étude porte sur une variable quantitative en fonction de deux (ou plus) variables qualitatives. Dans un modèle croisé, les variations observées peuvent venir d'un facteur, de l'autre, ou de la combinaison des deux.

Elle examine l'influence conjointe de deux facteurs et si leur interaction a un effet significatif sur la variable dépendante.

* H0 : Les moyennes des niveaux des deux facteurs et leur interaction sont égales (absence de différence significative).
* H1 : Au moins une moyenne est significativement différente, ou il y a une interaction significative.

L'analyse de variance à deux facteurs implique la variation totale, résiduelle, factorielle 1, factorielle 2 et d'interaction. Les moyennes pour le facteur 1et 2 sont calculées, puis la variable de Fisher est utilisée pour effectuer des tests de conformité.

Trois statistiques de Fisher sont obtenues, une pour les différences entre le facteur 1, une pour les différences entre le facteur 2, et une pour les différences dues à l'interaction des deux facteurs.

**Une image contenant texte, Police, jaune, capture d’écran

Description générée automatiquement**

Une image contenant texte, Police, capture d’écran

Description générée automatiquement En faisait le summary(lm()) dans l’Anova 2, on se retrouve, comme lors de la régression corrélation, avec des test de conformités sur base des αi issus du lm() avec H0 : µ = 0 pour l’intercept, suivi de H0 : αi = 0.

Ce qui revient à la réalité à un test d’homogénéité de moyenne telle que H0 : µi = µ1

Nous avions donc nos tests, nos hypothèses, et Fisher nous donnait la variable aléatoire :

****

Voilà qui explique pourquoi notre test H0 : µ1 = µ2 = µ3 s’appelle l’analyse de la variance, car le début de notre raisonnement part sur des écarts, les écarts des valeurs à leurs moyennes, ces écarts étant les ingrédients pour construire des écarts-types (qui nous mène à des estimations de variances => analyse de variance).

**13. ANOVA2:**

**Expliquer comment réaliser une étude de type ANOVA 2 à l'aide de R. Comment en arrive-ton aux définitions des coefficients et aux résultats statistiques présentés par R ?**

Une image contenant texte, diagramme, ligne, capture d’écran

Description générée automatiquement

**Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement**

**Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement**

**14. ANOVA:**

**Expliquer les différents types d’Anova 2 . Quand les utilise-t-on?**

**Une image contenant texte, Appareils électroniques, capture d’écran, Police

Description générée automatiquement**

**15. Termes:**

**Termes à définir brièvement, remettre dans son contexte, acronyme de … , utilité(s), rôle(s), exemple d’application :**

1. **Modèle croisé**

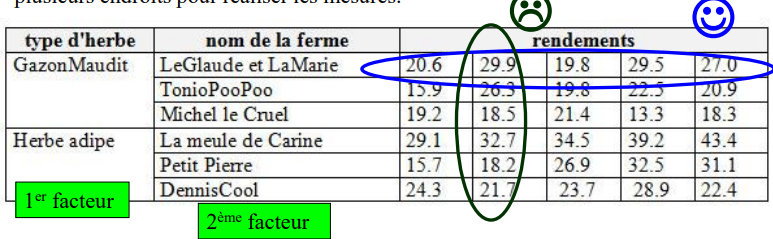
En Anova 2, modèle comprenant 2 critères de classification, chaque critère est sur le même pied que l’autre. On recherche les différences individuelles entre chaque sujets.

**Exemple :** On étudie le poids des porcs en fonction de l’éleveur et du type de porc élevé.

1. **Modèle hiérarchisé**

En Anova 2, modèle comprenant 2 critères de classification mais cette fois-ci l’un est subordonné à l’autre. On veut examiner les facteurs à différents niveaux de hiérarchie.

**Exemple :** Comparaison de rendements pour 2 type d’herbes et pour chaque on prend 3 prairies différentes pour tester les rendements mais il n’y a aucun lien entre les prairies et les 2 types d’herbes



1. **Interaction**

En Anova 2, c’est l’effet simultané de deux facteurs entrant en collision. Apparait dans l’analyse de variance dans un modèle croisé, il permet de voir si les effets sont dépendants d’un facteur ou non.

**Exemple :** Plusieurs ne possède pas toutes les mêmes types d’herbes variant ainsi l’écosystème et la qualité alimentaire des animaux.

1. **Contrastes**

En Anova 2, ce sont des coefficients pour tester une hypothèse plus spécifique qu’une simple égalité ou différence des moyennes.

**Exemple :** On a 2 médicaments visant au traitement pour la même maladie, et on va comparer leurs différents effets en changeant le sujet de l’expérience (homme, femme, enfant, jeune, vieux).

1. **Distance de Cook**

Dans la régression corrélation multiple, c’est un complément. Dans celui-ci on se demande si un point de notre nuage ne fausserait pas les régresseurs et ainsi pouvoir les identifier et les enlever. Pour ça on va créer des sous-modèles avec et sans ces points qui potentiellement biaises le résultat et on les compares au plan principale.

1. **Critère AIC**

Dans la régression corrélation multiple, c’est un critère de comparaison de modèles, qui estime la perte d’informations lorsque l’on utilise un modèle avec certains paramètres pour prédire les données en recherchant une perte minimale.

1. **Single linkage**

En Classification Ascendante Hiérarchique (CAH), Lorsque deux groupes sont à fusionner, la distance entre ces groupes est définie comme la plus petite distance(minimale) entre deux individus, l'un de chaque groupe.

**Exemple** : Supposons que les distances entre les individus A, B, C et D soient respectivement 1, 3, 2, 5. La distance minimale entre les groupes {A, B} et {C, D} serait de 2, car C (du premier groupe) et D (du deuxième groupe) ont la distance minimale.

1. **Complete linkage**

En Classification Ascendante Hiérarchique (CAH), La distance entre deux groupes est définie comme la plus grande(maximale) distance entre deux individus, l'un de chaque groupe.

**Exemple :** En reprenant les distances de l'exemple précédent, la distance maximale entre les groupes {A, B} et {C, D} serait de 5, car D (du deuxième groupe) et A (du premier groupe) ont la distance maximale.

1. **variabilité intra-classe**

Dans une CAH,quand on parle de partition, on parle que dans une même classe, d’individus forts proches les uns des autres. On parle de variabilité intra-classe petite lorsque l’on cherche une "bonne partition", par rapport à tout les individus de ce même groupe.

On calcule leur distance par rapport à la moyenne de chaque individu dans le groupe.

1. **variabilité inter-classe**

Dans une CAH,quand on parle de partition, on parle que dans une même classe, d’individus forts proches les uns des autres. On parle de variabilité inter-classe grande lorsque l’on cherche une "bonne partition". Permet savoir la distance entre 2 groupes, on compare leur écart entre leur 2 moyennes par rapport à la moyenne globale.

1. **Parangon (dans le cadre d’une HCPC)**

C’est l’individu représentant au mieux le groupe auxquels il appartient. C’est l’individu le plus proche du centre de gravité de la classe.