|  |  |
| --- | --- |
| M:\ADMIN\henallux_montgolfiere.png  Implantation IESN | Business intelligence  IG3 — C. Charlier |

# REGRESSION LINEAIRE SIMPLE

Ref : « exploration des données – Méthodes et modèles du data mining » Daniel T. Larose.  
<http://spss.espaceweb.usherbrooke.ca/pages/stat-inferentielles/regression-simple.php>  
site de Frédéric Bertrand, maître de conférences, université de Strasbourg (+ Myriam Maumy)  
« Statistiques pour l’économie et la gestion » Anderson-Sweeney-Williams-Camm-Cochran Ed. de boeck

## Introduction :

Alliance Data Systems (ADS) fournit des moyens de traitement des transactions, des services de crédit et des services de marketing à ses clients dans le domaine de la gestion des relations client, aujourd’hui en croissance. Les clients de ADS sont concentrés dans quatre secteurs : le commerce de détail, les stations-service les services publics et les transports . En 1983, Alliance a commencé à proposer des services de traitement des crédits aux entreprises appartenant aux secteurs du commerce de détail (y compris les stations-service) et de la restauration : cette société emploie aujourd’hui plus de 6500 personnes et offre ses services à des clients à travers le monde. Gérant de plus de 140000 points de vente aux Etats-Unis, ADS traite plus de 2,5 milliards de transactions par an. La société se place au deuxième rang des sociétés américaines privées de services de crédit. En 2001, ADS a fait une première offre publique d’achat et est maintenant cotée à la bourse de New-York.

L’un des services marketing d’ADS consiste à élaborer des campagnes promotionnelles par courrier. Grâce à sa base de données contenant des informations sur les habitudes d’achat de plus de 100 millions de consommateurs, ADS peut cibler les consommateurs qui seront les plus sensibles à une campagne promotionnelle. Le bureau de développement analytique utilise les méthodes de régression pour construire des modèles permettant de mesurer et de prévoir la sensibilité des consommateurs à des campagnes marketing ciblées. Certains modèles de régression prédisent la probabilité d’achat des individus recevant une réduction, d’autres prédisent le montant dépensé par les consommateurs qui effectuent un achat.

Lors d’une campagne promotionnelle particulière, une chaine de magasins souhaitait attirer de nouveaux consommateurs. Pour prévoir l’effet de la campagne, les analystes de ADS ont sélectionné un échantillon de consommateurs dans leur base de données, ont envoyé à ces individus des bons d’achat et ont ensuite collecté des données sur les transactions de ces clients : le montant ‘achat ainsi que plusieurs variables spécifiques à chaque consommateur susceptibles d’être utiles pour prévoir les ventes. La variable spécifique à chaque consommateur la plus pertinente pour prévoir le montant des achats, était le montant total des dépense s effectuées dans des magasins similaires au cours des 39 derniers mois. Les analystes d’ADS ont effectué une régression entre le montant des achats et le montant dépensé dans des magasins similaires : où correspond au montant des achats et le montant estimé des achats et x le montant dépensé dans des magasins similaires.

En utilisant cette équation, nous pouvons prédire qu’une personne qui a dépensé 10000 dollars au cours des 39 derniers mois dans des magasins similaires, dépensera 47,2 dollars en réponse à la campagne promotionnelle ciblée. Dans ce chapitre, vous apprendrez à effectuer ce type de régression.

NB : Le modèle final développe par les analystes de AS incluait également plusieurs autres variables, augmentant ainsi le pouvoir prédictif de l’équation précédente, telles que la possession ou non d’une carte de crédit bancaire, le revenu estimé et le montant moyen dépensé par visite dans un magasin particulier. Dans le chapitre suivant, nous verrons comment de telles variables additionnelles peuvent être incorporées dans un modèle de régression multiple.

## Objectifs

Les décisions prises par un responsable sont souvent basées sur la relation qui existe entre deux ou plusieurs variables. Par exemple, après avoir considéré la relation entre les dépenses publicitaires et les ventes, un responsable marketing peut essayer de prévoir les ventes pour un montant donné de dépenses publicitaires. Autre exemple, un fournisseur d’électricité peut se servir de la relation entre la température journalière maximale et la demande d’électricité pour prévoir la demande en électricité, en se basant sur les températures journalières maximales prévues pour le mois suivant. Parfois, un responsable peut se fier à son intuition pour déterminer le type de relation qui lie deux variables. Cependant, s’il est possible d’obtenir des données, une procédure statistique, appelée analyse de la régression, permet de construire une équation indiquant de quelle manière les variables sont liées.

Dans la terminologie utilisée dans le cadre d’une analyse de la régression, la variable que l’on cherche à prévoir, notée y, est la variable dépendante ou variable cible ou variable expliquée. La variable ou les variables utilisées pour prévoir la valeur de la variable dépendant sont appelées variables indépendantes ou variables explicatives.

Décider : de réformer le matériel car on a mis en relation le temps d’utilisation (y) et l’usure (x). Ainsi, il est possible d’anticiper la décision.

Prévoir : la consommation d’énergie nécessaire pour une production prévue (si on a mis en relation la production (x) et la consommation d’énergie (y)).

En maintenance, prévoir un modèle qui prédit la durée de vie d’un composant sur le site où il se trouve : on peut déterminer lors de petites défectuosités, des caractéristiques (des facteurs) qui influencent celles-ci et donc prédire la suite du fonctionnement.

Contrôler : exemple : dans les tunnels où on utilise de l’explosif pour creuser, on peut mettre en relation l’avancement des travaux du tunnel et la consommation d’explosifs. On peut utiliser cette relation pour voir si on n’a pas volé des explosifs.

## PRINCIPE DE LA REGRESSION DES MOINDRES CARRES:

Nous disposons de n points (x1,y1),(x2,y2),…,(xn,yn) et nous souhaitons trouver la « meilleure » droite.   
Comment trouver la « meilleure droite » ?

Voici le nuage de points :

Parmi toutes les droites qui approchent un ensemble donné de points, celle vérifiant la propriété suivante :  
 est minimum   
est la meilleure droite d’ajustement.

Ceci revient au même de minimiser f(a,b)=

Une droite présentant cette propriété est dite s’ajuster aux données au sens des moindres carrés et est appelée droite de régression des moindres carrés. Après démonstration, on obtient :

Donc et

L’équation estimée de la régression linéaire simple :

b s’appelle l’intercepte. C’est l’endroit sur l’axe des y où la droite de régression croise l’axe des y, ce qui donne la valeur estimée de la variable cible quand la variable prédictive est égale à 0. Dans certaines situations de régression, une valeur 0 pour la variable prédictive n’a pas de sens. Par exemple, si la variable cible est le poids des élèves avec comme variable prédictive la taille, l’image de la valeur 0 de la variable prédictive n’a pas d’interprétation. Par contre, dans l’exemple « céréales » (sucre – taux nutritionnel), une valeur de 0 pour la teneur en sucre a du sens et, dans ce cas, le taux nutritionnel estimé pour une céréale ayant une teneur en sucre de 0 est de 59,3.

a est la pente de la droite de régression. Cette pente indique le changement estimé dans la variable cible par unité d’accroissement de x. Dans l’exemple « céréales », nous interprèterons a = -2.4008 comme suit : pour tout accroissement de 1gr dans la teneur en sucre, le taux nutritionnel estimé décroît de 2,4008 points.

## PARAMETRES DE LA REGRESSION et INFERENCE

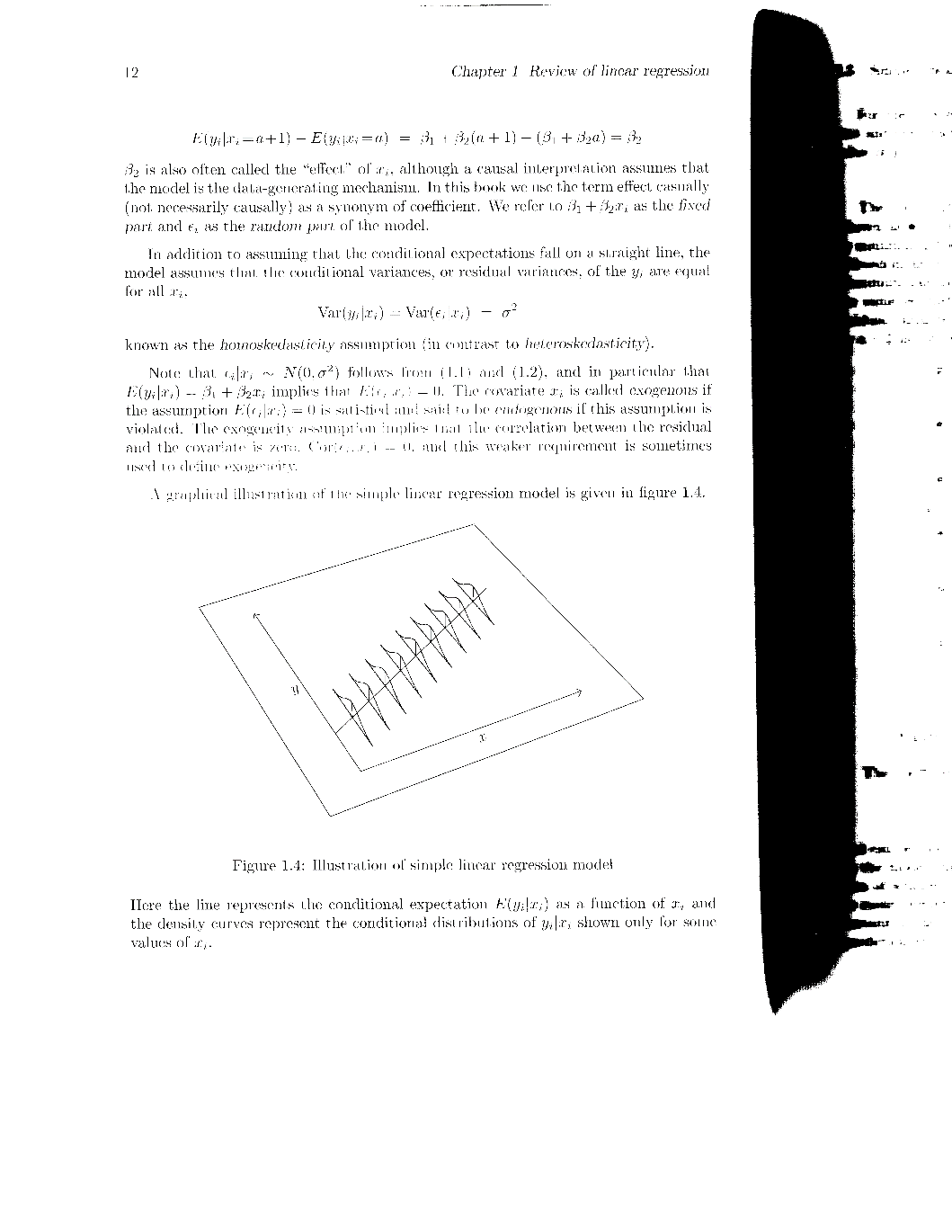
Bien sûr, une droite de régression des moindres carrés pourra toujours être trouvée pour estimer la relation entre deux variables continues mais cela ne garantit pas que cette régression soit utile. La question est « comment peut-on déterminer si une équation de régression estimée particulière est utile pour faire des prévisions » ?

Dans l’exemple « sucre – taux nutritionnel», si l’idée est de réaliser une régression linéaire sur les quelques valeurs données, nous nous arrêtons là. Si l’idée est que ces valeurs représentent un échantillon pour créer un modèle qui lie la taux nutritionnel d’une céréale au sucre qu’elle contient de manière générale alors la lecture sera :

nous possédons un échantillon de taille N et nous cherchons un modèle linéaire de type dans lequel

* le meilleur estimateur de est
* le meilleur estimateur de est
* est une variable qui représente le comportement individuel

Si nous avions accès à toute l ‘information, nous pourrions constater qu’une même abscisse correspond à différentes valeurs de .

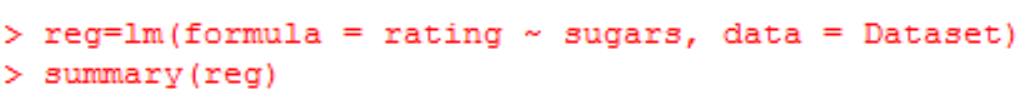


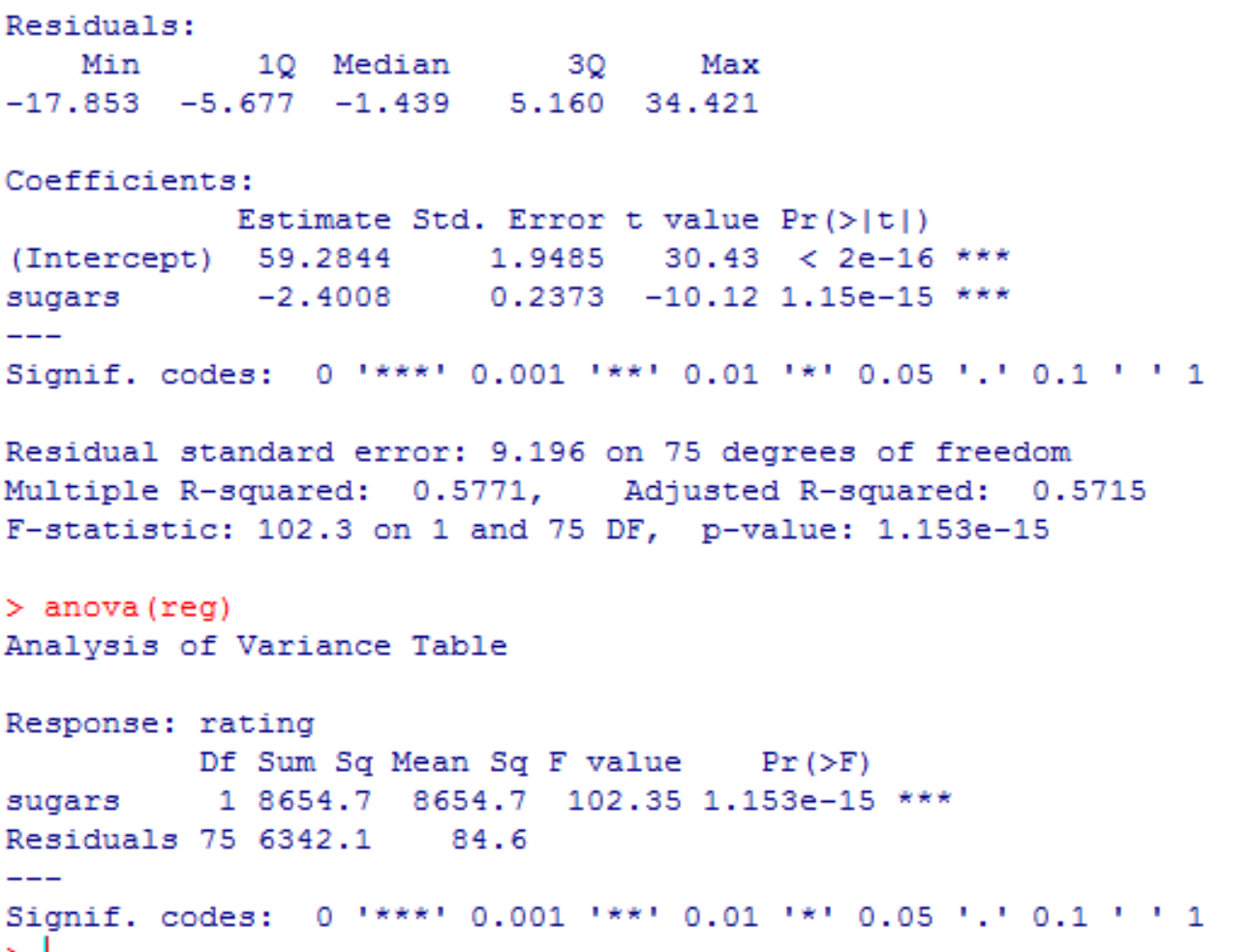
Ainsi, lorsque est fixé, est aléatoire et la composante aléatoire de est le .

Plusieurs droites de régression existent. La droite des moindres carrés consiste à minimiser, en principe, mais vu que nous ne les connaissons pas, cette somme sera estimée par = (cfr ci-dessus). Les estimateurs de sont respectivement et . Nous noterons donc la valeur estimée correspondant à .

Les logiciels statistiques donnent quelques indicateurs permettant de se faire une idée sur la pertinence de la régression et l’utilité du modèle.

### Analyse de la régression dans R





### Coefficient de détermination.

L’indicateur statistique sert à mesurer la qualité de l’ajustement de la régression càd mesure comment l’approximation linéaire s’ajuste réellement aux données observées.

Calcul :

* Si représente la valeur estimée de la variable cible, alors représente l’erreur de prévision ou le résidu,   
  SSE (som of squares error) =SCE (somme des carrés des erreurs) = représente une mesure globale de l’erreur dans la prévision résultant de l’usage de l’équation de régression estimée.   
  Dans notre exemple « céréales », SSE=6342,1. Est-ce beaucoup ?

* Supposons que nous n’ayons pas accès à la variable sucre, le meilleur estimateur pour les y serait , la moyenne de l’échantillon. Clairement, notre estimateur du taux nutritionnel serait dégradé dans l’ensemble puisque la droite d’estimation aurait pour équation y=42,558. Sur le schéma ci-dessous, nous voyons que les observations du départ semblent plus groupées autour de la droite de régression que de la droite de y=, ce qui suggère que l’erreur globale prédite est plus faible quand nous utilisons l’information sucre. Cela indique qu’utiliser l’information de la variable explicative améliore nos estimateurs du taux nutritionnel.  
  Nous définissons SST (som of squares total) = SCT (somme des carrés totale) = . La SST mesure la variabilité totale dans les valeurs de la seule variable cible. Remarquons SST=(n-1)\*Var(y).
* Nous aimerions obtenir une mesure du degré dont l’équation de la régression estimée améliore les estimateurs . L’indicateur statistique SSR (som of square regression) est une mesure de l’amélioration globale de la fiabilité de prévision en utilisant la régression par opposition à ignorer l’information de la variable explicative SSR =. Nous avons SST=SSR+SSE. SSR mesure la portion de la variabilité dans la variable cible y qui est prise en compte par la régression linéaire. Ici, la variable sucre explique 8654,7 dans le total de la variabilité à expliquer.
* En résumé : variabilité totale sur Y = variabilité de Y expliquée par la régression + variabilité de Y non expliquée par la régression.
* Nous avons mesure la qualité de l’ajustement de la régression. Ici vaut 0,5771.

### Coefficient de corrélation

r = = - 0,76 est un indicateur de la force de la relation linéaire entre les deux variables qualitatives. Ses valeurs peuvent être comprises entre -1 et 1.

Interprétation :

* les valeurs de r proches de 1 indiquent que les variables sont positivement corrélées càd quand la valeur x s’accroit, les valeurs y ont tendance à s’accroitre également (r > 0,7 variables corrélées ; si r est compris entre 0,33 et 0,7 légère corrélation)
* les valeurs de r proches de -1 indiquent que les variables sont négativement corrélées càd quand la valeur x s’accroit, les valeurs y ont tendance à diminuer (idem -0,7 et -0,33)
* les autres valeurs de r indiquent que les variables ne sont pas corrélées càd quand x s’accroit, les valeurs de y ne sont pas affectées.

### Ecart-type de l’estimateur s.

L’indicateur s est important à prendre en compte, c’est l’écart-type de la régression. , moyenne de la somme des erreurs au carré mesure la variabilité de la variable cible laissée non expliquée par la régression.   
NB : m est utilisé lors de la régression multiple, il représente le nombre de variables explicatives.

s = = 9,196 mesure la fiabilité des estimateurs produits par la régression. C’est l’estimateur du résidu typique, la différence typique entre la valeur prédite (estimée) et la valeur réelle. s représente la précision des prévisions générées par l’équation de la régression estimée. Plus la valeur de s est faible mieux c’est. Que veut dire une valeur faible de s ?

### Table ANOVA

En résumé, la table ANOVA présente :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source des variations | Somme des carrés | ddl | Moyenne des carrés | F |
| Régression | SSR | m | MSR=SSR/m | MSR/MSE |
| Erreur (résidus) | SSE | n-m-1 | MSE=SSE/(n-m-1) |  |
| Total | SST | n-1 |  |  |

### Exercice :

Calculez tous les paramètres de la régression tentant d’expliquer le taux nutritionnel par la quantité de fibres présentes dans la céréale – à partir d’excel et à partir de R. Interprétez tous ces résultats et comparez cette régression à la précédente.

### Test F

Il existe plusieurs démarches pour tester la validité de la linéarité d’une régression simple. Ces différents tests sont équivalents et reviennent à faire le test du coefficient de corrélation linéaire contre

Le test F dont les résultats sont donnés par l’ANOVA consiste à voir si la variance expliquée par la régression est significativement plus élevée que la variance non expliquée.

De manière plus précise, le test F réalisé ici est le suivant :

Test : (càd la relation linéaire est non significative) (càd relation linéaire significative)

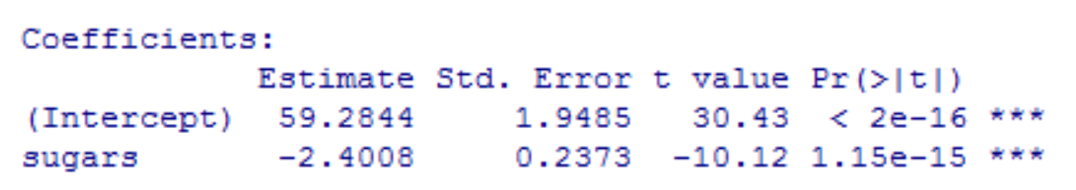
Fobs = la p-valeur est plus petite que

En supposant H0 vraie, nous avions maximum une probabilité 10(-15) d’obtenir un tel échantillon. Ainsi, nous considérons avoir suffisamment de preuve pour rejeter H0 et affirmer que le modèle de régression linéaire est significatif.

### Test t

Pour rappel, le test t a pour but de déterminer si la valeur d’espérance d’une population de distribution normale et d’écart-type inconnu est égale à une valeur déterminée . Pour ce faire, nous tirons de cette population un échantillon de taille n dont on calcule la moyenne et d’écart-type s.

Ici, l’hypothèse nulle teste contre en utilisant la loi de student à (n-2) degrés de liberté



Dans l’exemple ci-dessus :

P-valeur = .

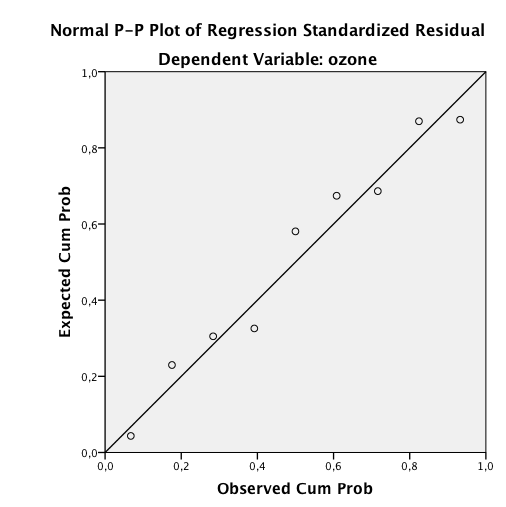
Le procédé est le même pour tester

### Hypothèses de régression.

La régression des moindres carrés est une méthodologie puissante et élégante. Cependant, pour éviter les erreurs, il y a lieu de vérifier certaines hypothèses :

Le modèle de régression où représente les termes d’erreur qui sont modélisés par des variables aléatoire indépendantes et identiquement distribuées N(0, Par conséquent,

* les sont des variables aléatoires indépendantes normalement distribuées N(
* est une variable aléatoire normale car ce paramètre dépend des variables qui sont des variables aléatoires normales
* est une variable aléatoire car ce paramètre dépend de qui est une variable aléatoire normale.

NB : Si nous voulons utiliser l’analyse de régression d’une manière strictement descriptive avec aucune inférence et pas de construction de modèle, nous n’avons pas besoin de nous inquiéter au sujet de l’hypothèse de validation. Cependant, si nous souhaitons faire de l’inférence et construire un modèle, les hypothèses doivent être vérifiée

Pour vérifier les hypothèses concernant les résidus, les logiciels fournissent des graphiques comme celui ci-dessus. Si le nuage de points de la probabilité normale ne montre aucune déviation systématique de la linéarité et si le nuage de points des résidus ne montre aucune caractéristique discernable, nous pouvons conclure qu’il n’y a pas de preuve graphique de violation des hypothèses de la régression et nous pouvons alors procéder à l’analyse de la régression.

## INDIVIDUS HORS NORMES – POINTS DE LEVIER ELEVES - OBSERVATIONS INFLUENTES

### Un individu aberrant :

Un individu aberrant est une observation qui a un résidu très important en valeur absolue.

Les deux observations avec les résidus absolus les plus importants sont « All bran with extra fiber » et « 100% bran ». Pour ces deux observations, nous remarquons que l’estimation est la plus éloignée de l’observation.

Il est préférable d’utiliser les résidus standardisés pour identifier les individus aberrants :

(cfr ci-dessous)

Une règle empirique est de marquer les observations dont les résidus standardisés excèdent 2 en valeur absolue comme valeurs aberrantes.

### Point de levier élevé.

Un point de levier élevé est une observation qui est extrême dans l’espace des variables prédictives. Le levier pour la ième observation se calcule comme suit : . Plus les observations en x sont éloignées de la moyenne, plus le levier est élevé. Un levier supérieur à environ ou peut être considéré comme un levier élevé.

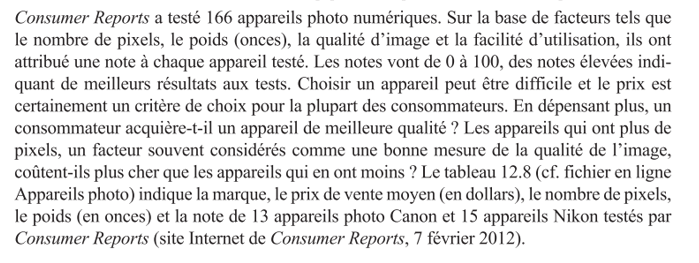
### Observations influentes.

Une observation est influente si sa présence ou son absence change significativement le cours des choses. Un individu aberrant peut ne pas être influent . De même, un point de levier élevé peut ne pas être influent. En général, les observations influentes combinent les caractéristiques d’un résidu important et d’un point de levier élevé. Cependant, il est possible pour une observation de ne pas être vraiment marquée ni comme individu aberrant, ni comme levier élevé mais d’être quand même influente grâce à la combinaison des deux caractéristiques.

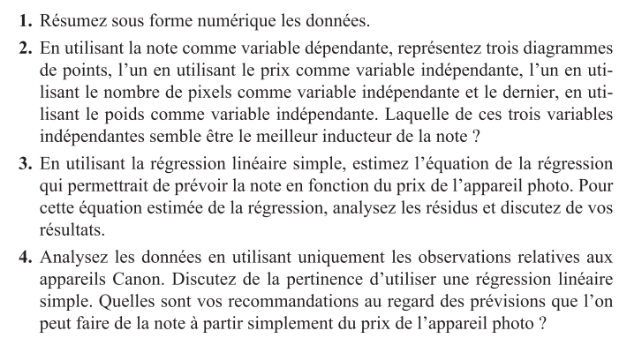
La distance de Cook mesure le niveau d’influence d’une observation en prenant en compte à la fois la taille du résidu et le montant du levier pour cette observation. La distance de Cook de la ième observation prend la forme suivante :

Une règle empirique pour déterminer si une observation est influente est que la distance de Cook excède 1.

## Problème 1



Cfr fichier AppareilsPhoto.xlsx

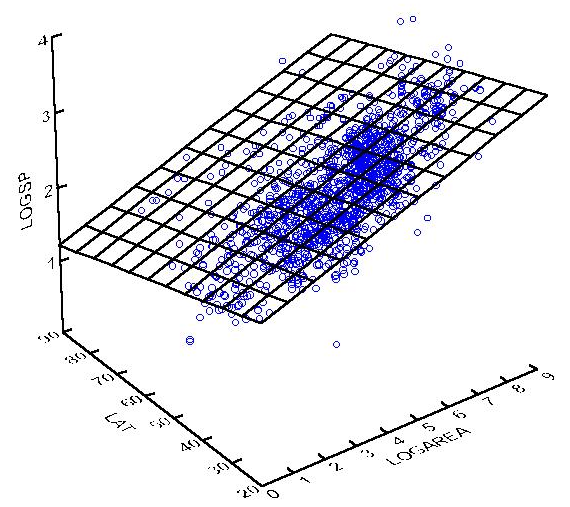


# REGRESSION LINEAIRE MULTIPLE

Ref : « exploration des données – Méthodes et modèles du data mining » Daniel T. Larose.  
Ecouter le cours Coursera Andrew G. « Machine learning » Université de Stanford

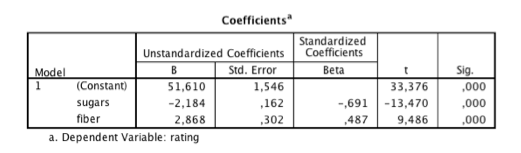
Généralement, les analystes de données sont intéressés par la relation entre la variable cible et un ensemble de (plus d’une) variables prédictives. La plupart des applications de data mining tirent profit de la profusion des données, avec certains ensembles de données incluant des centaines ou des milliers de variables, beaucoup d’entre elles pouvant avoir une relation linéaire avec la variable cible. La modélisation de la régression multiple fournit une méthode élégante pour décrire de telles relations.

Un modèle de régression multiple utilise une surface linéaire telle que le plan ou l’hyper plan pour approximer la relation entre la variable réponse continue et un ensemble de variables prédictives (éventuellement catégoriques).

*image : http://ordination.okstate.edu/MULTIPLE.htm*

L’équation de la régression estimée avec deux variables prédictives s’écrit En général avec m variables prédictives :

Voici le résultat des coefficients :

(3.1)

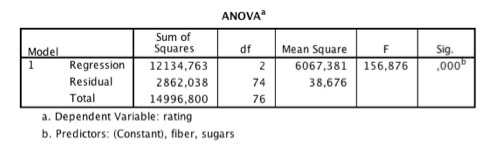
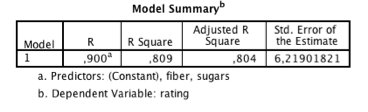
Il faut interpréter 1 = -2,184 comme la décroissance estimée du taux nutritionnel pour l’accroissement d’une unité du contenu en sucre quand le contenu en fibre est maintenu constant. Même principe pour interpréter 2.

Les niveaux élevés en fibres semblent être associés avec les taux nutritionnels élevés et les niveaux élevés de sucre semblent être associés à des taux nutritionnels faibles.

Le résidu est représenté par la distance verticale entre les données et le plan ou l’hyper plan de la régression.   
Exemple : le résidu lié au blé Spoon Size Shreddes : 0 gr de sucre, 3gr de fibres et un taux nutritionnel de 72,8018. La droite de régression aurait donné : Le résidu =

Chaque observation a son résidu. A partir de ces résidus, nous pouvons calculer une mesure globale des erreurs estimées : SSE = . De même, la SSR = et la SST= se calculent comme dans la régression linéaire simple.

La table ANOVA (ci-dessous) est présentée de la même manière.

R2 est le coefficient de détermination multiple et représente la proportion de variabilité dans la variable cible qui est prise en compte par sa relation linéaire avec l’ensemble des variables prédictives.

Chaque fois qu’une nouvelle variable prédictive est ajoutée, le coefficient de détermination augmente. Il augmente de façon significative si la nouvelle variable prédictive est utile sinon il augmente mais pas de façon significative.

Pour avoir une idée plus précise de l’intérêt de la nouvelle variable, nous définissons Raj = 1-(1-Cette formule permet de pénaliser les modèles qui incluent des variables prédictives inutiles.

L’erreur typique dans l’estimation est fournie par l’écart-type de l’estimation s, ici s=6,219 gr. Lorsque la variable prédictive ajoutée est utile, la valeur de s décroît et a tendance à augmenter lorsque la nouvelle variable n’est pas utile. Ceci fait de la variable s un meilleur indicateur de l’utililité d’une nouvelle variable que R2.

## MODELE DE REGRESSION MULTIPLE

où

* représentent les paramètres du modèle. Ceux-ci sont des constantes dont la valeur véritable reste inconnue, qui est estimée à partir des données en utilisant les estimateurs des moindres carrés ;
* représente le terme d’erreur . Ces variables aléatoires doivent être normales, indépendantes, de moyenne nulle et de variance constante.

Ceci implique que, pour y :

Les valeurs de la variable réponse sont des variables aléatoires normales et indépendantes de moyenne   
 et de variance .

## INFERENCE DANS LA REGRESSION MULTIPLE

### Test t pour la relation linéaire entre x et y

Ici, l’hypothèse nulle teste contre . Ceci revient à tester le modèle  
 contre .

La seule différence entre les deux modèles est la présence ou l’absence du terme en i. Ainsi, l’interprétation des résultats pour ce test t doit inclure le fait que les autres variables prédictives sont maintenues constantes.

Sous l’hypothèse nulle, le test statistique suit une distribution en t avec n-m-1 degrés de liberté où se réfère à l’écart-type pour la pente de la ième variable prédictive. Le test est réalisé pour le coefficient de chaque variable prédictive et pour le terme indépendant (cfr tableau 3.1). Dans ce tableau, les tests sont significatifs, nous rejetons l’hypothèse nulle donc il y a une preuve d’une relation linéaire entre le taux nutritionnel et le contenu en sucre (fibre) en la présence de contenu en fibre (sucre). Nous pouvons dire aussi que la contribution marginale de la variable « fibre » lorsqu’elle est introduite après la variable « sucre » est significative avec une p-valeur inférieure à 0,000..

### Test F pour la significativité du modèle de régression dans son ensemble.

Les test F considère la relation linéaire entre la variable cible y et l’ensemble des variables prédictives prises comme un tout.

Les hypothèses sont :

H0 :

H1 : Au moins un des n’est pas égal à 0.

Ainsi, le modèle sous H0 est .

Nous devons rejeter l’hypothèse nulle quand la valeur de l’indicateur statistique du test F est importante. Cet indicateur suit une distribution C’est un test unilatéral avec p=P(). S’il y a rejet de H0, cela signifie que le modèle global de régression est significatif.

## La multicolinéarité.

(ref : « Exploration des données » D.T.Larose)

Dans les modèles mis au point, il est nécessaire de se garantir contre la multicolinéarité, une condition où certaines variables prédictives sont corrélées les unes aux autres. La multicolinéarité conduit à une instabilité dans l’espace des solutions qui peut mener parfois à des résultats incohérents. Par exemple, dans un ensemble de données avec une sévère multicolinéarité, il est possible que le test F pour la régression globale soit significatif alors qu ‘aucun test t ne l’est.

L’idée de la régression multiple est de construire un plan (ou un hyperplan) qui approxime les observations au sens des moindres carrés. Prenons le cas de deux variables explicatives ( auquel cas c’est un plan qui est recherché. Si celui-ci est construit sur la base de deux variables (mathématiquement vecteurs) non corrélées alors sont orthogonales ou indépendantes. Dans un tel cas, les variables prédictives forment une base solide sur laquelle la surface de réponse peut rester robuste, ce qui fournit des estimateurs des coefficients chacun avec une faible variabilité

La forte variabilité associée avec les estimateurs signifie que différents échantillons peuvent produire des estimateurs des coefficients avec des valeurs largement différentes selon l’échantillon sur base duquel le modèle est produit. Même si une telle instabilité est évitée, l’inclusion de variables qui sont fortement corrélées tend à surestimer un composant particulier du modèle puisque le composant est essentiellement compté en double.

Pour éviter la multicolinéarité, l’analyste devrait étudier la nature des corrélations parmi les variables prédictives ( en ignorant temporairement la variable cible) en générant la matrice des coefficients de corrélation linéaire et en analysant les facteurs d’inflation de la variance :

Il faut demander que les facteurs d’inflation de la variance (VIF) soient rapportés : (facteur d’inflation de la variance pour . Cet indicateur mesure la corrélation entre le ième variable prédictive et le reste des variables prédictives.

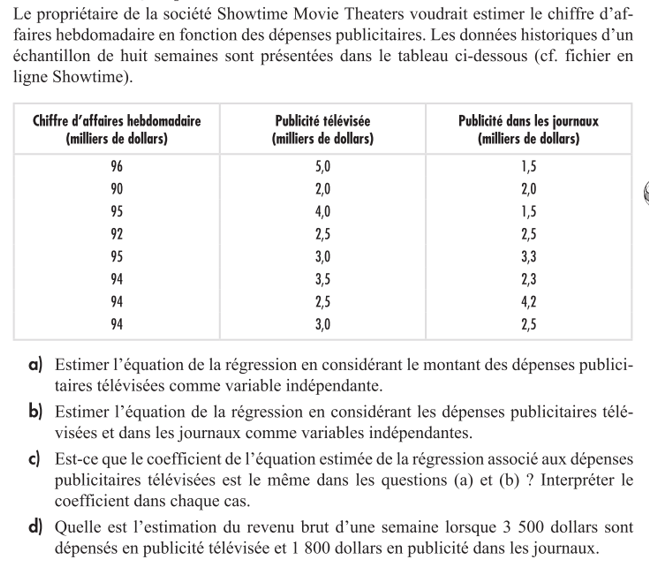
Comportement du VIF : Si est complètement corrélée avec les variables prédictives restantes alors donc . Il n’y a cependant pas de limite supérieure au VIF car, quand le degré de corrélation entre et les autres variables prédictives s’accroît, le s’accroît aussi.

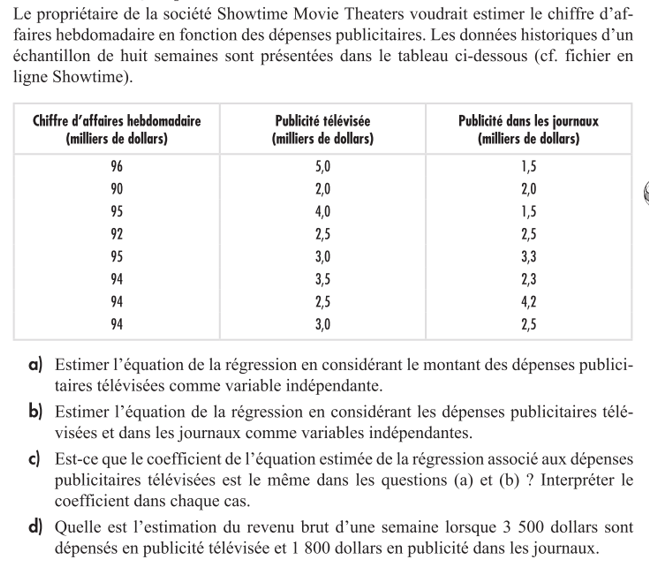
Quel effet ces changements dans VIF ont sur ? Nous avons .   
Si la variable est fortement corrélée aux autres produira une inflation de l’erreur standard . Augmenter les estimateurs de la variance résultera en une dégradation de la précision de l’estimation.

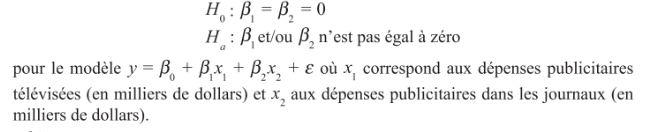
Règle :   
-  : indicateur de colinéarité modérée (  
-  :indicateur de colinéarité sévère (

Dans le chapître suivant, nous allons expliquer une manière de sélectionner des variables indépendantes.

## Exercices





1. A partir des paramètres SCT et SCR donnés par R, calculez R² et R² ajusté ;
2. Lorsque, seules les dépenses publicitaires télévisées sont considérées comme variables indépendantes, R²=0,653 et R²ajusté=0,595. Les résultats de la régression sont-ils préférables ? Justifiez.
3. Testez les hypothèses suivantes avec  :  
   
4. Utilisez pour tester la significativité de Qu’en concluez-vous ?   
   Idem pour

# ANALYSE EN COMPOSANTES PRINCIPALES (ACP).

(ref : WikiStat : « Introduction à l’analyse en composantes principales »//  « Analyse en composantes principales »  
« Exploration des données » D.T . Larose)

## Introduction

Comment généraliser le nuage de points tracé dans le cas de deux variables et aborder la structure de corrélation présente entre plus de deux variables ? En appliquant l’analyse en composantes principales.

Mathématiquement, l’analyse en composantes principales est un simple changement de base : passer d’une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations.

Prenons l’ensemble de données ci-dessous qui indique les cotes (entre 0 et 20) obtenues par 9 étudiants dans 4 disciplines.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Math | Physique | Franc | Anglais |
| Jean | 6 | 6 | 5 | 5,5 |
| Alain | 8 | 8 | 8 | 8 |
| Annie | 6 | 7 | 11 | 9,5 |
| Monique | 14,5 | 14,5 | 15,5 | 15 |
| Didier | 14 | 14 | 12 | 12,5 |
| André | 11 | 10 | 5,5 | 7 |
| Pierre | 5,5 | 7 | 14 | 11,5 |
| Brice | 13 | 12,5 | 8,5 | 9,5 |
| Evelyne | 9 | 9,5 | 12,5 | 12 |

## Représentation matricielle.

Soient p (=4) variables statistiques réelles observées sur n (=9) individus.   
Ainsi, sur le ième individu. Ces observations sont regroupées dans une matrice X d’ordre n.

Dans l’exemple :

* A chaque individu i est associé le vecteur contenant la ième ligne de X. C’est un élément d’un espace vectoriel noté E de dimension p. E est appelé espace des individus.
* A chaque variable est associé le vecteur contenant la jème colonne centrée. C’est un élément d’un espace vectoriel F de dimension n. F est appelé l’espace des variables.
* Lorsque les variables sont centrées et représentées par des vecteurs de F :
  + la longueur d’un vecteur représente un écart-type
  + le cosinus de l’angle entre deux vecteurs représente une corrélation.

## Décomposition spectrale de la matrice des covariances.

Exercice 1 : Etude des 4 variables séparément.

Analysez séparément ces 4 variables : graphiquement et numériquement.

Exercice 2 : Etude des variables 2 à 2

Analysez les liaisons entre les variables 2 à 2.

Pour l’étude simultanée des 4 variables, les choses se compliquent car chaque individu est représenté dans un espace de dimension 4. Les objectifs de l’ACP sont

* de revenir à un espace de dimension réduite q < p(par exemple 2) en déformant le moins possible la réalité. Nous voulons donc obtenir le résumé le plus pertinent des données initiales;
* de représenter les individus de manière « optimale » en minimisant les déformations du nuage des points, dans un sous-espace Eq de dimension q < p ;
* de représenter graphiquement les variables dans un sous-espace Fq en explicitant « au mieux » les liaisons initiales entre ces variables ;
* de faciliter l’utilisation de techniques comme la régression linéaire en fournissant des variables orthogonales.

Remarquons que, dans notre exemple, les données sont toutes du même ordre de grandeur. Si les variables ne sont pas homogènes, elles doivent être préalablement réduites.

Etape 1 :

Réaliser une standardisation des données.

Etape 2 :

Calculer la matrice symétrique des variances-covariances notée où mesure le degré avec lequel les deux variables varient ensemble. Une covariance positive indique que quand une variable s’accroît, l’autre tend à s’accroitre. Une covariance négative indique que quand une des variables s’accroit, l’autre a tendance à décroitre

Les composantes principales représentent un nouveau système de coordonnées, trouvé en faisant tourner le système original le long des directions de variabilité maximum. Remarque : si sont indépendantes alors leur covariance est nulle.

NB : la covariance n’est pas redimensionnée donc changer les unités de mesure change la covariance, ce qui n’est pas le cas avec le coefficient de corrélation linéaire : .

Etape 3 :

Matrice des corrélations : .

Lorsque les données sont standardisées, la matrice des variances-covariances et la matrice des corrélations sont les mêmes.

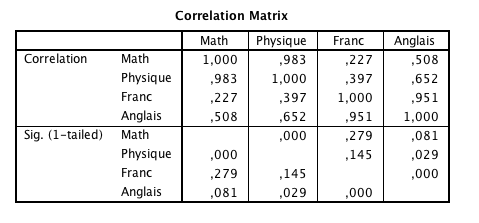
Etape 4 :

La i-ième composante principale de la matrice des données standardisées Z=[Z1,Z2,…,Zn] est donnée par   
 où est le i-ième vecteur propre de la matrice des corrélations.   
Les composantes principales sont les combinaisons linéaires des variables standardisées de Z telles que les variances des soient les plus élevées possible et que les ne soient pas corrélées.

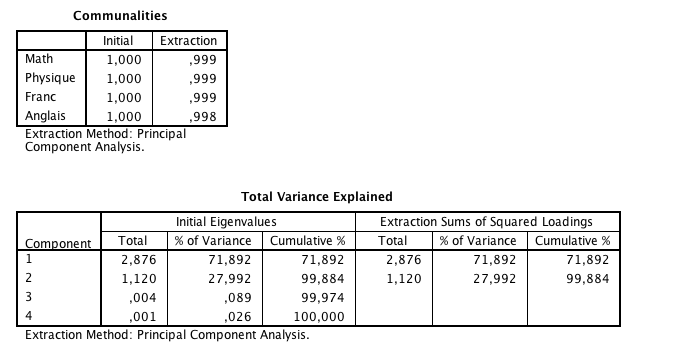
Revenons à notre exemple :

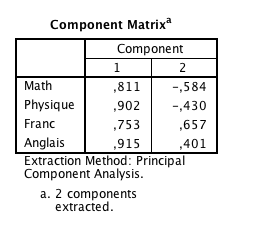
En réalisant une ACP sur la matrice des cotes, nous obtenons les résultats suivants :

Matrice des corrélations :



Cette matrice représente les coefficients de corrélation linéaire entre les différentes variables. Bien sûr, elle est composée de 1 sur sa diagonale (en effet, une variable est toujours 100% corrélée à elle-même) et elle est symétrique (en effet, le coefficient de corrélation linéaire entre la variable j et la variable i est le même que le coefficient de corrélation linéaire entre la variable i et la variable j). La valeur 0,983 pour les variables math-physique indique une forte corrélation (linéaire) positive entre les cotes obtenues en math et les cotes obtenues en physique par les différents individus. La valeur 0,227 pour les variables math-français indique une très faible corrélation positive entre les cotes obtenues en math et les cotes obtenues en français par les différents individus.



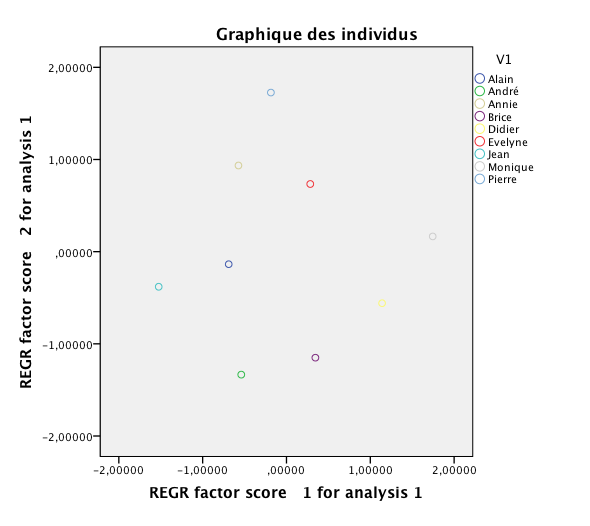


Le tableau ci-dessus montre que deux facteurs (deux vecteurs propres) ont été sélectionnés pour représenter les quatre variables simultanément. Ces facteurs sont des combinaisons linéaires des variables de départ.   
Axe 1 :  
Le premier est calculé (sur les données standardisées) en prenant 0,811\*cote de math + 0,902\*cote de physique +… . Ainsi, l’axe 1 représente une sorte de « note globale de l’étudiant ». Plus l’individu (cfr graphique des individus) est représenté « loin » vers l’axe 1 positif, plus sa note globale est bonne (exemple : Monique). Si l’individu est représenté vers le 0 de l’axe 1, il a une note globale plutôt moyenne (exemple : Evelyne). Plus l’individu est réprésenté « loin » dans l’axe 1 négatif, plus cela signifie que sa note globale est faible (exemple : Jean).

Axe 2 :

Le deuxième est calculé (sur les données standardisées) en prenant - 0,584 \* cote de math - 0,43 \* cote de physique + 0,657 \* cote de français + 0,401 \* cote d’anglais. Ainsi, l’axe 2 représente une sorte « d’écart entre les matières littéraires et les matières scientifiques ». Plus l’individu (cfr graphique des individus) est représenté « loin » vers l’axe 2 positif, plus l’écart entre ses notes dans les branches littéraires et le branches scientifiques est grand et ceci, au profit des branches littéraires (exemple : Pierre). Si l’individu est représenté vers le 0 de l’axe 2, ses notes sont alors assez équivalentes dans les branches scientifiques et dans les branches littéraires (exemple : Brice). Plus l’individu est réprésenté « loin » dans l’axe 2 négatif, plus cela signifie que l’écart entre ses notes dans les branches littéraires et le branches scientifiques est grand et ceci, au profit des branches scientifiques (exemple : Pierre).

En interprétant selon les deux axes, nous pouvons dire que Annie a des notes faibles, inférieures à la moyenne mais qu ‘elle a de meilleures notes dans les branches littéraires que dans les branches scientifiques.



Que peut-on dire des facteurs choisis ?

Le tableau « communalities » signale que 99,9% de l’information contenue dans les cotes de math est prise en compte en prenant ces deux facteurs. Idem pour les autres variables. Remarquons que ces valeurs sont souvent plus faibles quand même et qu’elles sont souvent différentes d’une variable à une autre, ce qui n’apparaît pas dans cet exemple. Le graphique des variables (components plot) indique cette information dans la longueur des vecteurs représentés. Il faut imaginer une cercle de centre « croisement des axes » et de rayon 1. Plus la longueur du vecteur représenté s’approche de 1, mieux il est représenté dans la nouvelle configuration. Puisqu’ici, 99,9% de l’information de chaque variable a été extraite, les quatre vecteurs (variables de départ) sont de même longueur et de longueur proche de 1. Si l’information contenue dans l’une de ces variables avait été extraite à 50%, la variable correpondante aurait été représentée sur une longueur 0 ,5.

Le tableau « total variance explained » présente les valeurs propres. En faisant le rapport entre la valeur propre du facteur i et la somme des valeurs propres, nous obtenons le pourcentage de variance expliquée par le facteur i. Exemple première valeur propre 2,876 /(2,876+1,12+0,004+0,001)= 0,7189 donc 71,892% de la variance est expliquée par le facteur 1 (cfr deuxième colonne). Nous constatons aussi, à la lecture de ce tableau (% cumulative variance) que 99,884% de la variance est expliquée par les deux facteurs.

Le graphique ci-dessous permet aussi de connaître la corrélation entre les différentes variables. Celle-ci est représentée par le cosinus de l’angle entre les deux vecteurs. Plus l’angle est petit (vers 0), plus le cosinus de l’angle tend vers 1 donc plus la corrélation est forte et positive (exemple : entre math et physique). Plus l’angle tend vers 90°, plus le cosinus de l’angle tend vers 0 donc plus la corrélation est faible (exemple : entre math et français). Si l’angle se situe entre 90° et 180°, il indique une corrélation négative allant respectivement de très faible (nulle à 90°) à très forte.

