Pr. Staccini

Biostatistiques

MODÈLES MULTIVARIÉS

Dulclaudiax

Sommaire

***Disclaimer** : ++*
Coucou mes bébouchats,
Dernier cours de biostat de l'année :((jrigole j'ai encore des fiches complètes à sortir lol) sur les modèles multivariés !
Petit aparté pour vous dire que cette fiche résume TOUT le cours du prof. Il dit que ce qui tombera à l'examen ce sera uniquement la régression linéaire simple et la régression logistique (donc pas la régression linéaire multiple apparament), mais l'an dernier on a eu des petits items sur le reste du cours, donc je vous ai tout mis au cas où vous voudriez lire l'ensemble une fois pour éviter les mauvaises surprises le jour-J...
Il dit aussi qu'il ne demandera pas les formules (ouf)
Sur ce, bonne lecture :)

DÉFINITIONS

La statistique

Méthode qui consiste à observer et étudier **une ou plusieurs propriétés communes** chez un groupe d'êtres, de choses ou d'entités.

Une statistique

Un **nombre calculé à partir d'une population** *(d'êtres, de choses ou d'entités)*.

Population

Collection (*d'êtres, de choses ou d'entités)* ayant des **propriétés communes**. Ce terme est hérité d'une des premières applications de la statistique : la démographie.
Exemple : un ensemble de parcelles de terrain étudiées, une population d'animaux, un groupe de patients présentant une maladie définie, l'ensemble des plantes d'une espèce donnée, une population d'humains habitants un lieu particulier…

Individu

Élément de la population.
Exemple : un patient, un insecte, une plante…

Variable

Propriété commune aux individus que l'on souhaite étudier. Elle peut être :

Qualitative

Ex : appréciation de la parcelle, l'état de santé de l'insecte, couleur des pétales, appartenance religieuse

Quantitative continue

= Variable numérique pouvant prendre n'importe quelle valeur réelle. *Ex : le taux d'acidité du sol, la longueur de l'insecte, la longueur de la tige, l'indice de masse corporelle (IMC)*

Quantitative discrète

= Variable numérique avec un saut minimum obligatoire entre deux valeurs successives (ex : nombres entiers). *Ex : la somme (sur tous les jours) du nombre de vaches présentes sur la parcelle, l'âge de l'insecte (en jours), le nombre de pétales sur la fleur, le nombre d'année d'études (réussies) depuis la petite école*

2 types de statistiques

Statistique descriptive	Statistique inférentielle
Son but est de **décrire**, c'est-à-dire résumer ou représenter par des statistiques les **données disponibles** quand elles sont nombreuses	Les données sont considérées **incomplètes** et elle a pour but de tenter de **retrouver l'information sur la population initiale**. La prémisse est que chaque mesure est une variable aléatoire suivant la loi de probabilité de la population.
Questions types: • Représentation graphique • Paramètres de position et de dispersion • Diverses questions liées aux grands jeux de données	*Questions types :* • Estimation de paramètres • Intervalles de confiance • Tests d'hypothèses • Modélisation (ex : régression linéaire)

La statistique peut être...

→ Univariée

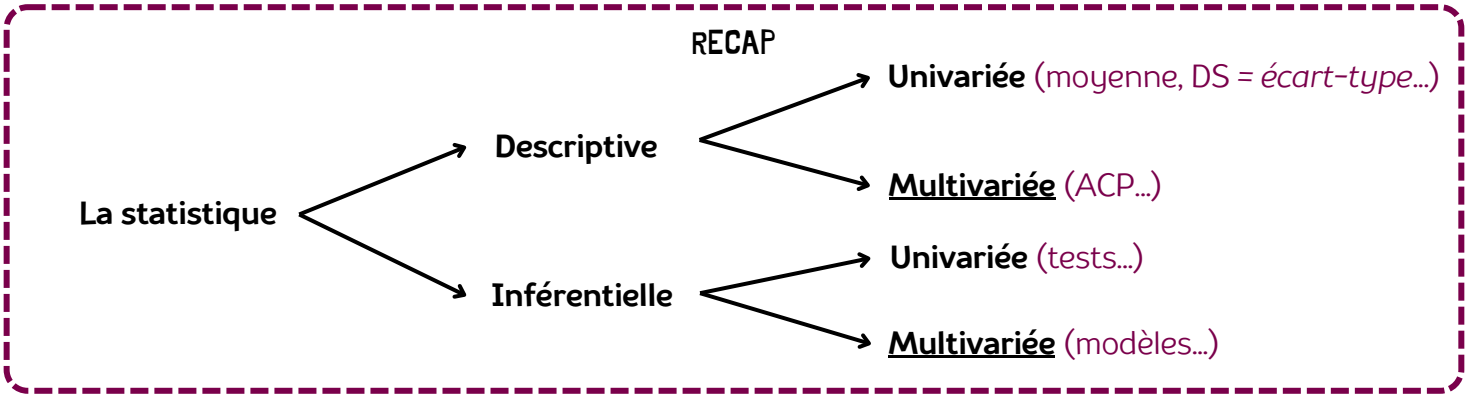
Il n'y a qu'**une seule** variable qui rentre en jeu.

→ Multivariée

Il y a **plusieurs variables** qui rentrent en ligne de compte.

- Deux variables entre elles : analyse **bivariée**
- Plusieurs variables : analyse **multivariée**

Le schéma est : il y a **une variable expliquée** et **plusieurs variables explicatives** indépendantes deux à deux.



RÉGRESSION LINÉAIRE

Point'tut

En statistique, la régression est une méthode permettant de proposer un modèle mathématique pour **expliquer les relations entre les observations**.
La **régression linéaire simple** consiste à proposer une **droite** pour expliquer une variable aléatoire **quantitative** par une autre.

Le **coefficient de corrélation linéaire** mesure la **liaison entre 2 variables aléatoires** (donc plus il est élevé, plus le lien entre ces variables est fort).

Les variables ont un rôle symétrique. Cependant, la question à résoudre peut être plus précise et libellée sous la forme suivante : *"Les valeurs prises par une variable Y dépendent-elles des valeurs de X ?"*
Ici, les deux variables ne sont pas considérées de manière équivalente :

- **Y** est la **variable à expliquer**, aussi appelée **variable dépendante**. Elle est la variable dont on veut expliquer les valeurs;
- **X** est la **variable explicative**, aussi appelée **variable indépendante**. Elle est la variable que l'on veut utiliser pour expliquer Y.

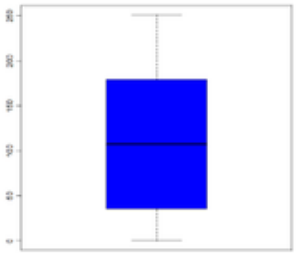
La courbe qui décrit les variations de Y en fonction de X s'appelle la **courbe de régression de Y en X.** On peut, en première approximation, chercher à assimiler cette courbe à une **droite**.

Exemple : lien entre taille et âge

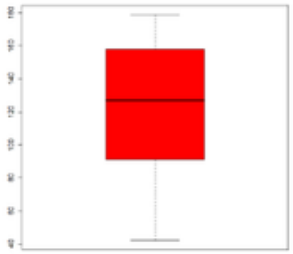
On souhaite étudier le lien entre la taille (en cm) et l'âge (en mois) des filles sur un échantillon de 637 filles.
On se pose plusieurs questions :

- Existe-t-il un lien entre la taille et l'âge ?
  - S'il n'existe pas de lien, on obtiendra une droite parallèle à l'axe des abscisses (toute variation de X ne produit aucune variation de Y).
- Quand l'âge augmente, est-ce que la taille augmente aussi ?
- Connaissant l'âge, peut-on prédire la taille ?
  - But médical : peut-on détecter des retards de croissance ? Ou au contraire, dans le cas de la médecine légale, lorsqu'on retrouve un os humain, peut-on déterminer l'âge et le sexe ?

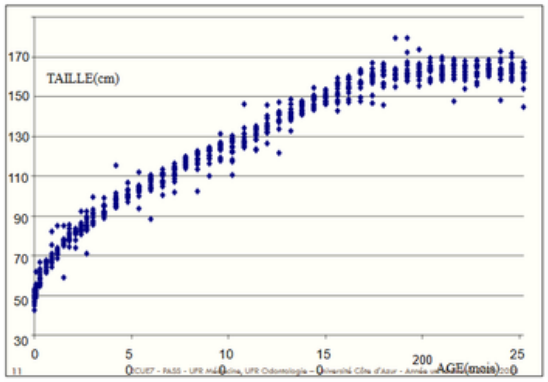
NB : on parle dans ce cours de **régression** (linéaire ou logistique) **simple ou multiple** → **simple** = 1 seule variable explicative; **multiple** = plusieurs variables explicatives.



m = 112,12 mois
s^2 = 6265,86 $mois^2$



m= 122,83 cm
s^2 = 1317,43 cm^2



Ce qu'on voit à gauche ce sont des diagrammes en boîte : ils permettent de visualiser la distribution d'une variable quantitative. En bleu on voit la variable représentant l'âge, et en rouge la variable représentant la taille. Le trait au milieu représente la médiane. Les boîtes indiquent où se trouvent les 50% centraux des données.

Ce qu'on voit à droite, c'est un nuage de points. Chaque point représente une fille, avec son âge en abscisse (X) et sa taille en ordonnée (Y). C'est donc une analyse bivariée descriptive.

La question qu'on se pose donc et qu'on va étudier est : **comment la taille évolue-t-elle en fonction de l'âge ?**

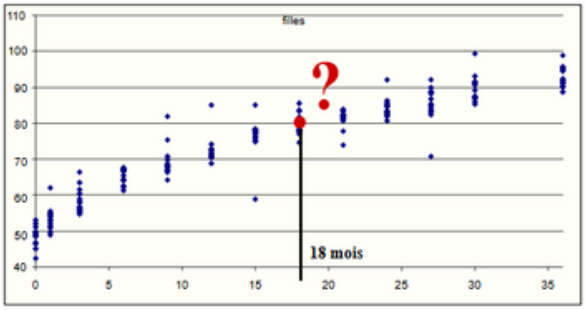
On pose **taille = f(âge)** *(= "la taille est fonction de l'âge")*
Autrement dit, on se demande, pour une variation de X, quelle est la variation de Y ?

Pour cela, on va donc utiliser la régression → on parle de **régression de Y en X** avec :

- Y = taille (cm)
- X = âge (mois)

On cherche donc à savoir comment évolue la taille en fonction de l'âge pour chaque valeur d'âge *(équation)* ou bien encore quelle est la taille pour un âge donné *(valeur et intervalle de confiance).*

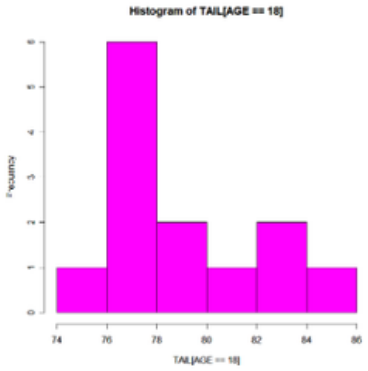
Exemple au sein d'un groupe de filles



Chez les filles de 18 mois, on va cherche la taille moyenne, la variance de la taille et la distribution.

Méthode pour déterminer l'âge à 18 mois :

- On stratifie les données.
- On sélectionne les filles de 18 mois.
- On calcule les paramètres de la distribution (moyenne et variance)
- On calcule un intervalle de confiance à 95% de la moyenne.



Résultats :

- Moyenne observée : M(T/A=18) = 79,23cm
- Variance observée = V(T/A=18) = 9,36cm^2

On parle de **distribution conditionnelle** → **valeur de la taille sachant l'âge**. (= T/A)

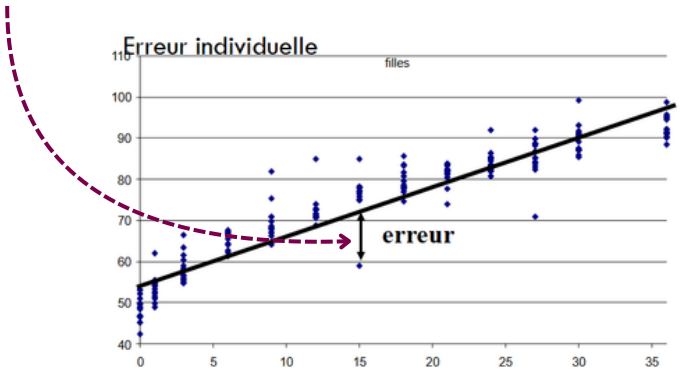
Fonction de régression

La taille en fonction de l'âge, aussi écrite Moyenne(Taille/Age) = f(Age), peut s'exprimer par une **fonction f** qui est une **droite affine** de type **y = ax + b**.

Dans notre exemple, la fonction affine est donc : **Espérance(Taille / Age) = α + β × Age.**

Pour chaque sujet, on définit que la Taille = α + β × Age + ε avec ε étant **l'erreur individuelle**.

L'erreur individuelle (**ε**) est **l'écart entre la valeur obtenue** par la fonction (y = ax + b) et la **vraie valeur observée.**



S'il n'existe **pas de lien** (= pas de corrélation) entre X et Y (ici l'âge et la taille), **alors toute variation de X n'entraine aucune variation de Y**.
Graphiquement, on obtient donc une **droite parallèle** à l'axe des abscisses d'équation y = constante.
Au niveau de l'équation y = ax + b, on aurait alors a = 0. *(donc y = b = constante)*

Les individus (les points) présents sur le graphique plus haut sont tout de même assez éloignés de la moyenne observée. On va donc essayer de minimiser ces écarts :

Pour chaque individu, par rapport à la moyenne, l'erreur est tant positive que négative. Pour **minimiser ces écarts** et s'affranchir du signe, il faut donc **les passer au carré**. On va donc faire ce qu'on appelle **la somme des carrés des écarts** (SCE). *(c'est cette méthode qu'on utilise pour obtenir notre droite des moindres carrés qu'on appelle aussi donc droite de régression !)*

La régression linéaire est le modèle le plus simple pour permettre :

- une **interprétation** (lien ou non entre les deux variables), permise par la valeur du **coefficient de régression** qui englobe dans son calcul la pente de la droite, donc la valeur de ß
- une **estimation de α et ß** pour que la droite d'ajustement minimise l'erreur individuelle
- la **prédiction** et l'**extrapolation**

La **droite d'ajustement**, aussi appelée **droite de régression**, permet donc de résumer au mieux le nuage de points.

Point'tut

Pour résumer *(parce qu'on se perd entre cours et exemple…)* : la **régression** c'est **prouver que l'une des deux variables permet de prédire l'autre**, c'est-à-dire montrer qu'à partir de X on peut prédire Y.

Petite anecdote : On appelle ça régression parce qu'au départ on avait observé un phénomène de retour vers la moyenne, et en anglais c'est "regression" qu'on utilise pour qualifier ce phénomène. Les valeurs très grandes ou très petites avaient tendance à se rapprocher de la moyenne. Le mot régression signifie simplement trouver une équation qui explique comment une variable Y dépend d'une variable X.

On essaie alors de trouver les valeurs de la droite d'équation : **$Y = \alpha + \beta X + \varepsilon$** avec

- **Y** la variable à expliquer
- **X** la variable explicative
- **α** l'ordonnée à l'origine (c'est la valeur de Y pour X=0)
- **β** la pente (c'est la variation moyenne de la valeur de Y pour une augmentation d'une unité de X)
- **ε** l'erreur aléatoire / l'erreur individuelle

Principe de l'estimation

Pour obtenir notre droite de régression, il faut donc trouver la bonne formule pour sa fonction. **On veut donc estimer α et ß tel que ε soit le plus petit possible.**

εi représente l'écart entre la droite et le point i.
Pour chaque valeur de X, on a yi = α + βxi + ε
Or, E(Y/X) = α + βX
Donc εi = yi – E(Y/X)
On calcule la somme des carrés des écarts : $SCE = \sum_{i=1}^{n} (\varepsilon i)^2$

On cherche à estimer α et β tel que la SCE soit **la plus petite possible**.

Point'tut

La **distance d'un point à la droite** est la distance verticale entre l'ordonnée du point observé et l'ordonnée du point correspondant sur la droite. Cette distance d'un point à la droite représente **l'erreur ε.**

Pour s'affranchir du signe de l'erreur ε, on calcule la **somme des carrés des distances de chaque point à la droite** (SCE). La **droite de régression** est alors **la droite qui minimise la somme des carrés des écarts** *(donc c'est la droite qui passe le plus proche de chaque point du nuage).*

Estimation de la pente β

$$\beta = \frac{cov\,(XY)}{var\,(X)}$$ avec :

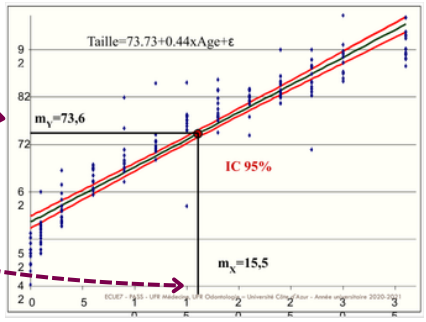
- **cov(X/Y)** = covariance de X et de Y *(la convariance indique dans quelles mesures deux variables varient ensemble)*
- **var(X)** = variance de X

Dans l'exemple, β = cov(TAILLE/AGE)/var(AGE) = 0,437703

Estimation de l'ordonnée à l'origine a

La droite passe par les points mY et mX.
On a mY = α + βmX donc $\alpha = mY - \beta mX$

Dans l'exemple, α = 73,729



L'équation finale s'écrit donc :

$$Y = \alpha + \beta X + \varepsilon \text{ ou } E(Y/X) = \alpha + \beta X$$

Dans notre exemple, on a ***Taille = 73,73 + 0,44 Age + ε*** *ou* ***E(Taille/Age) = 73,73 + 0,44 Age.***

Point'tut

Une particularité de la droite de régression est de **passer par le point moyen théorique de coordonnées (mx ; my),** où mx est la moyenne empirique de X et my est la moyenne empirique de Y sur l'échantillon.

L'estimation de l'ordonnée à l'origine α est déduit de la pente β et des coordonnées du point moyen (mx ;my) par la formule suivante : **α = mY – βmX**

Interprétation

Interprétation de la pente β

- **β = 0** : **pas de lien**, évolutions indépendantes
- **β < 0** : **évolution** en **sens contraire**
- **β > 0** : **évolution** dans le **même sens**

Interprétation de l'ordonnée à l'origine a

$$E(Y/X = 0) = \alpha$$

Test de la pente à 0

Si **β = 0**, alors il n'y a **pas de lien entre Y et X**.

Le lien entre Y et X est-il significatif ? Autrement dit, est-ce que ß≠0 ?

Soit b une estimation de ß, la fluctuation de b observée peut être due au hasard.

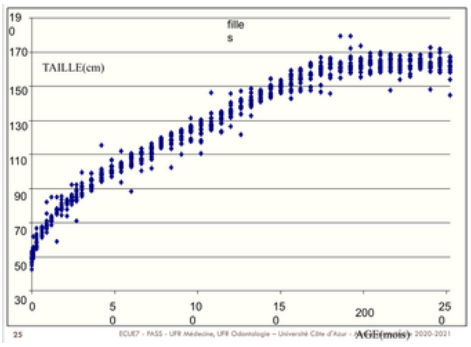
On note les hypothèses :

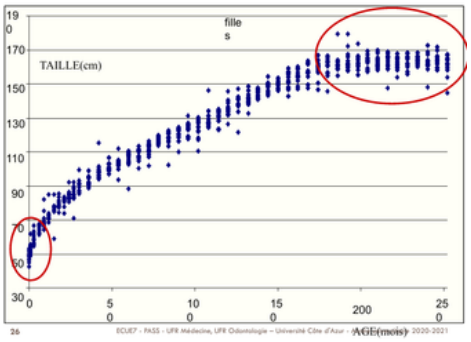
- **H0** : **β =0**, il n'y a pas de lien entre X et Y
- **H1** : **β ≠0**, il existe un lien entre X et Y

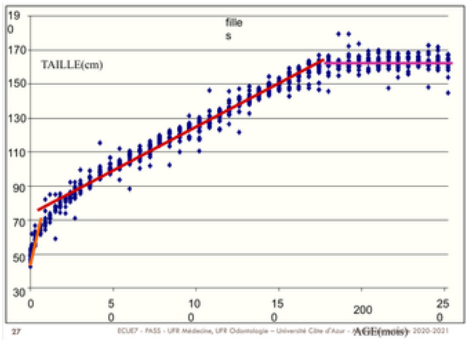
Sous H0, et si les conditions d'application sont respectées, on a une statistique $t_0 = \frac{b-\beta}{\sqrt{s_b^2}}$ une loi de Student à n-2 DDL, avec :

- L(Y/X) qui tend vers N
- V(Y/X) constante pour tout X
- à X donné, on a un Yi indépendant

→ **La régression est linéaire**.







Point'tut : le test de la pente

On veut appliquer un test statistique qui est le **test de la pente de la droite de régression**. La droite de régression d'équation **Y = α + βX comporte 2 paramètres** (α et β).

L'hypothèse nulle H0 est que la **pente β de la droite de régression de Y en X est égale à 0**, c'est-à-dire que Y est égal à α, ou encore que la droite de régression est horizontale et qu'il n'y a **pas de liaison entre X et Y**.

L'hypothèse alternative H1 est que **la pente β de la droite est différente de 0**.

Sous H0, le rapport de l'estimateur de la pente b sur son écart-type suit une **loi de Student à (n- 2) DDL**, où n est l'effectif de l'échantillon.

Le **test de la pente** consiste à **calculer la grandeur t_0** et à la **comparer à la valeur seuil t_α** sur la table de la loi de Student à (n-2) DDL.

Le hasard explique-t-il la fluctuation de b ?

Pour en juger, on va s'intéresser aux intervalles de confiance.

Intervalle de confiance de la pente

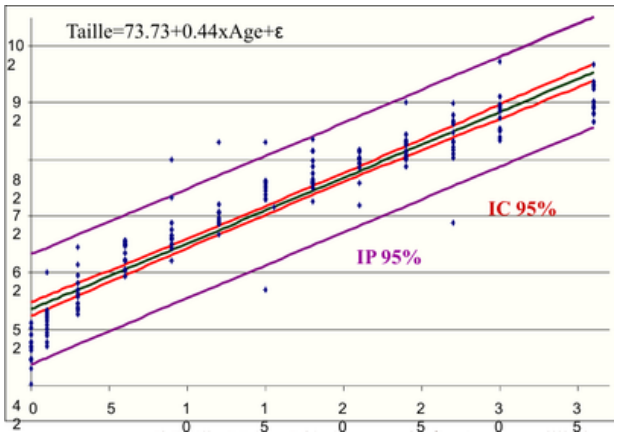
b tend vers t_{n-2} et on a $b \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{s_b^2}$

Si l'intervalle de confiance à 95% de b ne contient pas la valeur 0, dans ce cas, **b est différent de 0 au risque d'erreur 5%**.

Intervalle de confiance de la droite

$E(Y/X) = \alpha + \beta X$ estimé par $m_{y/x} = \alpha + bX$

d'où $m_{y/x} \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{s_{m^2_{y/x}}}$



Intervalle de prédiction

Pour un âge (X), on prédit la taille (Y) :

$Y_p = \alpha + bX$ soit **$Taille_p = 73{,}73 + 0{,}44Age$**

Précision de la prédiction

$y_p \pm t_{n-2,\frac{\alpha}{2}} \times \sqrt{{s_{y_p}}^2}$

Point'tut

Pour rappel, la régression linéaire est à la fois un outil de **prédiction** et un outil **d'inférence**, et le test de Student ainsi que les intervalles de confiance servent à faire cette inférence.

On retrouve les cours de statistiques descriptives et déductives car pour rappel, en régression linéaire, on **estime une droite sur un échantillon**, puis on utilise la **théorie statistique** pour savoir **si cette droite reflète un effet réel ou juste du hasard d'échantillonnage.**

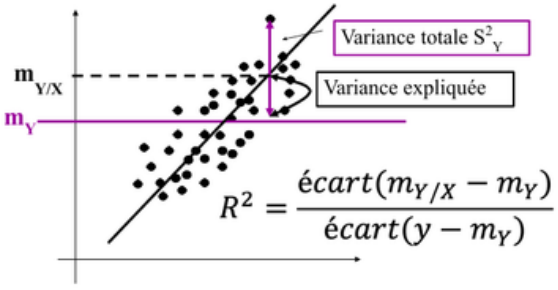
On se pose la question de l'**adéquation du modèle**, c'est-à-dire, **est ce que le modèle est un bon résumé des observations ?**

Pour cela, on va calculer le pourcentage de variance expliquée R^2 :

$$R^2 = \frac{part\,de\,la\,variance\,expliquee}{variance\,totale} = \frac{ecart\,(m_{y/x} - m_y)}{ecart\,(y - m_y)}$$

Variance totale : S_y^2

Pourcentage de variance expliquée :



$$R^2 = \frac{\sum (m_{y/x} - m_y)^2}{\sum (y - m_y)^2}$$

Exemple : R^2 = 88%

$\sqrt{R^2}$ = estimation du coefficient de corrélation entre X et Y

Point'tut : le pourcentage de variance expliquée

Sur la même lancée que les intervalles de confiance, le **pourcentage de variance expliquée R^2** nous permet de savoir **si notre modèle est bien représentatif de la réalité**. R^2 nous dit à quel point notre modèle colle aux données.

La valeur que prend R^2 correspond au **pourcentage de variance de Y qu'explique notre modèle** *(ex : le modèle explique 88% des variations de Y)*. On peut également dire que R^2 correspond à la variabilité totale de la variable à expliquer (Y) est due aux variables explicatives (X) *(ex : 88% des variations de Y sont expliquées par les variations de X).*

RÉGRESSION LOGISTIQUE

La **régression logistique** traite des cas lorsque la variable Y qu'on cherche à expliquer est qualitative, et plus précisément **binaire** (ex : malade oui/non).

On utilise ce modèle lorsque les conditions d'application de la régression linéaire ne sont pas remplies.

La (ou les) variable(s) explicative(s) X peuvent, quant à elle(s), être aussi bien quantitatives que qualitatives.

On note la fonction de notre variable : $Y = f(X1; X2; \ldots; Xn)$

Le but de la régression logistique est alors :

- D'**expliquer Y** en quantifiant l'association de Y pour chaque x_i
- De **prédire Y** à partir des nouvelles observations de x_i

Point'tut : la régression logistique

Contrairement à la **régression linéaire**, qui prédit une **valeur quantitative** *(par exemple, un poids, une tension artérielle ou un score)*, la **régression logistique** est une méthode statistique utilisée pour **prédire la probabilité d'un événement binaire**. Autrement dit, elle sert à **estimer la probabilité qu'un événement se produise ou non** *(par exemple: malade/non malade, succès/échec, vivant/décédé, etc.).*

Dans l'exemple qui va suivre, la régression logistique permet d'estimer la probabilité de décès en fonction de la dose d'un toxique ingérée.

Graphiquement, contrairement à la régression linéaire qui est représentée par une droite, la **régression logistique** suit une **courbe sigmoïde** (en forme de «S»), qui traduit le fait que la probabilité ne peut jamais être inférieure à 0 ni supérieure à 1.

En régression logistique, on interprète souvent les résultats à l'aide de **l'Odds Ratio (OR)**. L'OR indique combien les **odds** (ou «chances relatives») de survenue de l'événement sont **multipliées** lorsque la variable explicative X augmente d'une unité. C'est donc un indicateur de la **force** et du **sens** de l'association entre une variable et l'événement étudié.

- **OR = 1** : **pas d'association** entre X et l'événement (les chances restent identiques)
- **OR > 1** : **X augmente le risque**. *Ex : OR = 2 → les chances sont 2 fois plus élevées.*
- **OR < 1** : **X diminue le risque**. *Ex : OR = 0,5 → les chances sont réduites de moitié.*

L'éloignement de 1 mesure l'ampleur de l'effet (plus c'est extrême, plus l'effet est fort). *Dans l'exemple qui suit, si OR = 1,5 pour la dose de toxique → chaque unité supplémentaire augmente les chances de décès de 50%. Si OR = 0,8 pour un traitement → chaque unité de dose de traitement réduit les chances de décès de 20%.*

En résumé :

- **Régression linéaire** → prédiction d'une **valeur numérique** (variable quantitative)
- **Régression logistique** → prédiction d'une **probabilité** pour un **événement binaire**

Exemple : décès en fonction d'une dose de toxique

On cherche à savoir **comment varie la proportion de décès en fonction de la dose toxique ?**

Ici on voit que l'analyse est **bivariée**, et que les deux variables aléatoires en jeu sont d'une part **qualitative binaire** (décès oui/non → Y) et d'autre part **quantitative** (dose de toxique → X).

Pour créer notre modèle, on utilisera une fonction de la forme :

$$logit\,(p) = ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

Pour rappel, **l'estimation d'une probabilité est un rapport** *(par ex, la proba d'avoir une boule rouge parmi 3 boules rouges et 2 boules noires c'est le rapport 3/5)*

Pour pouvoir "transformer" un rapport en somme, on passe par la fonction logarithme :

$$log\,(A/B) = logA - logB$$

La fonction logit donne le log népérien de la côte d'un évènement, c'est-à-dire le rapport p/(1-p).

$$logit\,(p) = ln\left(\frac{p}{1-p}\right)$$

Point'tut : la fonction logit

La **fonction logit** est une transformation mathématique qui **convertit une probabilité p** (comprise entre 0 et 1) en un **nombre réel allant de −∞ à +∞**.

Pourquoi l'utilise-t-on ?

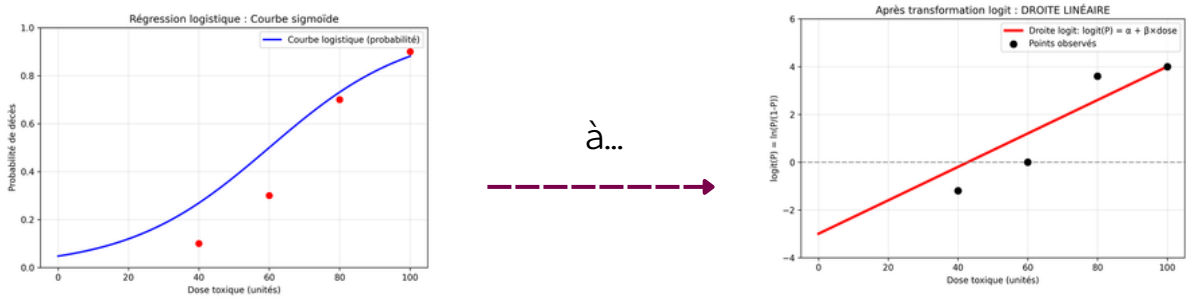
Comme la variable à expliquer (décès/vivant) est **binaire**, on travaille avec des **probabilités**. Mais les probabilités sont difficiles à manipuler mathématiquement (contraintes $0 \leqslant p \leqslant 1$). La fonction logit **linéarise** cette relation !

La fonction logit a pour formule :

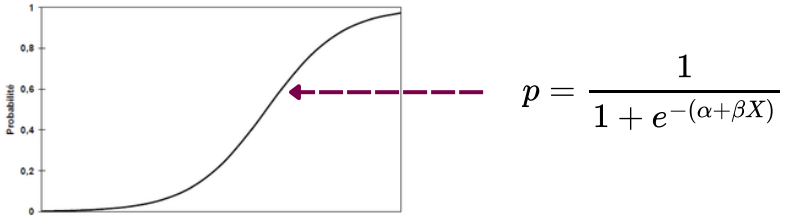
$$logit\,(p) = ln\left(\frac{p}{1-p}\right)$$

Graphiquement :

- Sans logit → **Courbe sigmoïde** (compliquée à modéliser)
- Avec logit → **Droite** (α + βX, facile à estimer !)



Après transformation logit, on obtient une **droite simple** qu'on peut analyser avec les méthodes de la **régression linéaire** classique !



$$p = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

Si la variable indépendante (variable explicative) est également binaire, alors pour calculer ce p et ce 1 - p on va procéder de la sorte :

	Chez les exposés	Chez les non exposés
Exposition	E = 1	E = 0
Probabilité d'être malade	$p_+ = p\left(M^+/E=1\right) = \frac{1}{1+e^{-(\alpha+\beta)}}$	$p_- = p\left(M^+/E=0\right) = \frac{1}{1+e^{-\alpha}}$
Probabilité de ne pas être malade	$1-p_+ = p\left(M^-/E=1\right) = \frac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$	$1-p_- = p\left(M^-/E=0\right) = \frac{e^{-\alpha}}{1+e^{-\alpha}}$

Odds Ratio (OR)

L'**Odds Ratio** (OR) ou rapport de côtes désigne la **force du lien entre les variables X et Y.** *Plus il est élevé, plus le lien est fort.*

Il est déterminé à partir de l'estimation des paramètres précédents :

$$OR = \frac{\frac{p_+}{(1-p_+)}}{\frac{p_-}{(1-p_-)}} = e^{\beta}$$

*Il est appelé "rapport de côtes" car c'est littéralement le **rapport entre deux côtes** : Odds Ratio = côte(exposé) / côte(non-exposé).*

Conditions d'application de la régression logistique

- Il doit y avoir une **relation linéaire entre logit(p) et X**

Pour rappel, la fonction logit(p) est : $logit\,(p) = ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$

- **Y** (la variable à expliquer/dépendante) doit être **binomiale ou multinomiale**

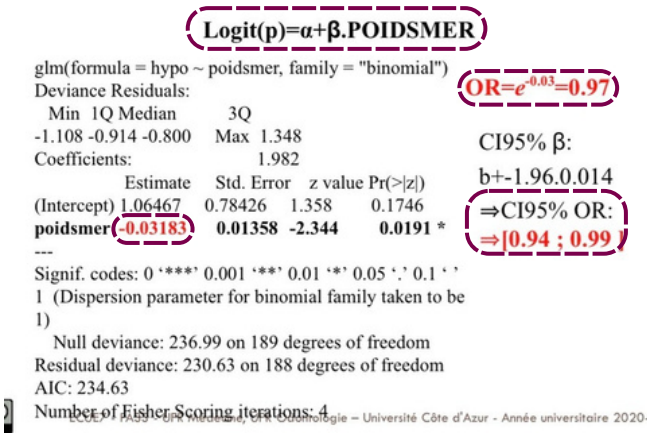
Binomiale = 2 issues possibles (ex : malade/ non malade), multinomiale = plusieurs issues possibles (ex : les 5 stades d'une maladie).

- Il doit y avoir un **codage "intelligent" des X catégoriels** pour interpréter les coefficients
- Il doit y avoir l'**indépendance des individus**

Exemple : facteurs d'hypotrophie à la naissance

Le poids de la mère est-il un facteur d'hypotrophie ?

Les informations importantes sont entourées



Interprétation:

p < 0,05 (le degré de signification, cf le cours stats déductives) donc on conclut que **l'OR** est significativement **différent de 1**, et donc qu'il **existe un lien significatif** entre le poids de la mère et l'hypotrophie dans le sens suivant : lorsque le poids de la mère augmente, le risque d'hypotrophie diminue.

Pour chaque unité de poids maternel, les chances d'hypotrophie sont **multipliées par 0,97** (= OR), autrement dit **diminuées de 3%** (1 - 0,97). On fait l'hypothèse d'un OR constant, quelque soit le poids maternel. Il s'agit d'une **relation linéaire** entre le risque d'hypotrophie et le poids maternel (dans l'espace logit). Si le modèle n'est pas une droite, alors on **modifie** le codage du poids maternel.

RÉGRESSION LINÉAIRE MULTIPLE

Pour reprendre l'exemple de la régression linéaire simple, on peut trouver **plusieurs causes** dans l'évolution de la taille Y :

- L'âge (X_1)
- Les facteurs socio-économiques (X_2)
- Les taux d'hormones de croissances (X_3) ...

Dans ce cas, on a : $E(Y/X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Estimation : α, β_1, β_2, β_3 sont estimés en tenant compte des 3 variables aléatoires X_1, X_2, X_3.

On parle alors **d'ajustement**

On peut envisager des **intéractions** :

$$E(Y/X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$$

Concernant ce qui suit, le prof a simplement énuméré les points sur son diapo, donc j'en ai déduit que c'était les différentes étapes nécessaires à la construction d'un modèle de régression multiple (que j'ai essayé de "traduire" au mieux).

- Tests des ß1, ß2, ß3 à 0 *(vérifier si chaque variable est utile)*
- Interprétation identique *(regarder l'effet quand les autres variables sont constantes)*
- Adéquation identique *(vérifier si le modèle explique bien (R^2))*
- Approche pas à pas *(optimiser en sélectionnant stepwise si nécessaire)*
- Choix des variables : notion de modèle *(choisir le modèle le plus simple (principe de parcimonie))*
- Variables très corrélées *(vérifier s'il y a des variables trop corrélées qu'il faudrait enlever)*

Exemple : prédire l'âge en fonction de 8 mesures

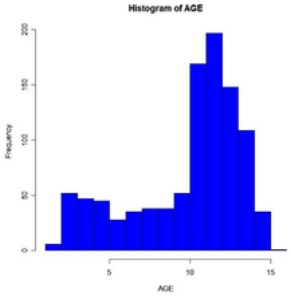
- Crâne (BIP)
- Tronc (LATHO)
- Membres supérieurs et inférieurs (LOMAIN, PERPOIGN, PERCHEV, PIEDS)
- Globales (STAT, POIDS)

On travaille sur un <u>échantillon de 1000 enfants de 2 à 16 ans</u>.

En moyenne, on trouve :

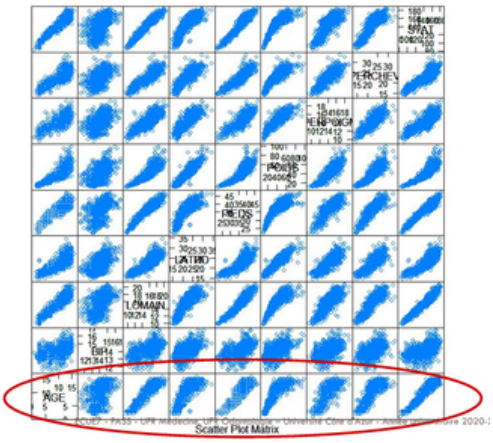
$$AGE = \alpha + \beta_1 \times BIP + \beta_2 \times LATHO + \beta_3 \times LOMAIN + \beta_4 \times PERPOIGN + \beta_5 \times PERCHEV + \beta_6 \times PIEDS + \beta_7 \times STAT + \beta_8 \times POIDS$$

Les statistiques descriptives nous indiquent :
mean(AGE) = 10,373 (=moyenne)
var(AGE) = 11,53541 (=variance)




```
Call: glm(formula = AGE ~ 1 + BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV +    PIEDS +
STAT + POIDS, family = gaussian)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.12658 -0.72416 -0.04954  0.67239
4.36643

Coefficients:     Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)-1.300e+01   8.684e-01  -14.966  < 2e-16 ***
BIP 3.312e-02           5.423e-02   0.611   0.54156
LATHO  1.219e-01        2.659e-02   4.583   5.17e-06 ***
LOMAIN 1.013e-01        5.947e-02   1.704   0.08877 .
PERPOIGN -1.370e-01     4.695e-02  -2.917   0.00361 **
PERCHEV                 2.597e-02  -1.792   0.07341 .
-4.654e-02
PIEDS   7.823e-04       2.612e-02   0.030   0.97611
STAT    1.546e-01       7.263e-03  21.289  < 2e-16 ***
POIDS  -2.047e-02       7.153e-03  -2.861   0.00431 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
AGE=-13+0,03BIP+0,1LATHO+0,01LOMAIN-0,14PERPOIGN-0,05PERCHEV
+0,001PIEDS+0,2STAT-0,02POIDS
AIC: 3010.6
Number of Fisher Scoring iterations: 2
```



A gauche, c'est une <u>matrice de nuages de points</u> (scatterplot matrix). Chaque ligne et chaque colonne représente une variable, ce qui signifie que chaque case représente le nuage de points entre 2 variables. La diagonale correspond aux histogrammes de chaque variable et ce qui est entouré en rouge correspond à la <u>variable à expliquer (ÂGE) vs toutes les variables explicatives.</u> Le but est de voir <u>quelles variables prédisent le mieux l'âge avant la régression multiple</u>.

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Que faut-il regarder ensuite ?
→ Les conditions d'application
→ Les intervalles de confiance des paramètres
→ L'adéquation : R^2

Adéquation: R^2 0.8989102

Point'tut : l'adéquation

L'adéquation (ou coefficient de détermination) R^2 est un autre nom pour le **pourcentage de variance expliquée** ! Il mesure la qualité d'ajustement d'un modèle de régression.

C'est la proportion de la variance totale de la variable dépendante expliquée par la/les variable(s) indépendante(s).

Sa valeur va de **0** (modèle nul) à **1** (ajustement parfait).

Dans cet exemple, $R^2 = 0.90$ → **environ 90% de la variable est expliquée par le modèle.**

On prouve la corrélation mais pas la causalité !

Sélection des variables du modèle

Guillaume d'Ockham (1285-1349) a posé la base de la sélection des variables du modèle avec le **principe de parcimonie** : **"les multiples ne doivent pas être utilisés sans nécessité"**. On n'ajoutera donc pas de nouvelles variables tant que celles présentes suffisent.

On cherche une **balance entre explication et prédiction**; s'il y a trop de variables alors notre modèle expliquera mieux mais perdra en prédiction. On parle aussi **d'overfitting** ou **d'hyperadéquation**.

Ainsi, la sélection de variables se fait **pas-à-pas (=stepwise)**:

- **Ascendant** = on ajoute les variables une à une
- **Descendant** = on retire les variables une à une
- **Double sens**

Pour sélectionner les variables, il existe un critère de sélection : le **score AIC (Akaike Information Criterion).**

Pour le calculer : $AIC = 2p - 2ln(L)$ avec p = nombre de paramètres et L= vraisemblance du modèle

On cherche le AIC le **plus petit possible**.

Dans l'exemple précédent :

AGE ~ 1 + STAT ; AIC = 3039.4

AGE ~ 1 + BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + PIEDS+ STAT + POIDS ; AIC = 3010.55

AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + STAT + POIDS ; AIC = 3008.55

RÉGRESSION LOGISTIQUE MULTIPLE

Exemple : L'hypotrophie à la naissance dépend-elle du tabagisme, de l'HTA, de l'âge maternel et du poids maternel ?

Dans ce cas de figure-là, il est nécessaire de faire attention aux **interactions** qu'il peut y avoir entre les variables, notamment ici entre l'HTA et le tabac et entre l'HTA et le poids.

On utilise **l'analyse univariée** grâce au test exact de Fisher, au test du Chi2 de Pearson, et au test t de Student pour l'HTA, le tabac, l'âge et le poids maternel.

Et on utilise des **tests d'interaction** (test exact de Fisher, test de Wilcoxon) pour l'étude des variables HTA.TABAC et HTA.POIDSMAT.

MÉTHODES PARTICULIÈRES

- Données de comptages : **régression de Poisson** (nombre d'évènements dans le temps)
- **Régression non linéaire**
- Données censurées (survie) :
  - **Estimation de Kaplan Meier ou Actuarielle**
  - **Test du Log Rank** (univarié), **Modèle de Cox** (multivarié)
- Séries temporelles (**Box-Jenkins**)
- Variabilité spatiale
- Analyse factorielle de données
  - **ACP, ACM, Arbres, CHA, Kmeans**...

Analyse en composantes principales (ACP)

Dans cette méthode de **réduction de dimension**, les variables sont **toutes quantitatives**. Les moyennes, variances, corrélations ont un sens.

On cherche à examiner la structure des données :

- Les individus se ressemblent-ils tous ?
- Existe-t-il des sous-groupes d'individus ?
- Des individus aberrants ?

On se demande **quelles sont les variables corrélées entre elles ?**
L'ACP permet d'interpréter facilement la matrice de corrélation.
Pour **p variables**, il existe **p(p+1)/2 corrélations possibles** !

Principes de l'ACP

Si les données ne comportaient que 2 variables : une représentation graphique suffirait pour répondre aux objectifs.
Mais, en général il y a **p variables** (on parle **d'espace à p dimensions**) et la représentation sous forme d'axes simples devient impossible.
⇒ L'idée est donc d'obtenir des **représentations approchées dans un espace en dimension 2** (2 dimensions) !

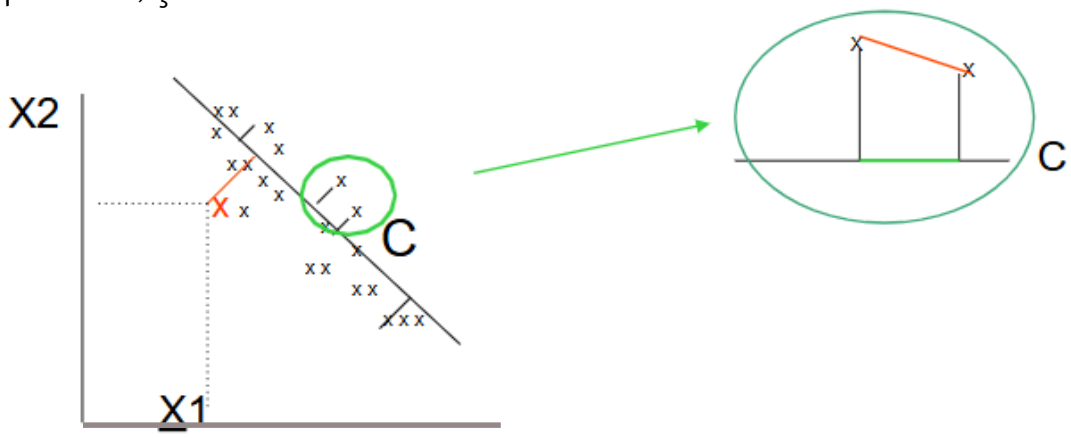
On estime qu'on a p variables, ce qui revient à parler d'une dimension p (Rp). Le but est d'obtenir des représentations en dimension 2 les plus fiables possibles.

Le critère sur lequel on va se baser va être la **conservation de la variance**, c'est-à-dire qu'on souhaite **conserver la distance entre les individus** lorsqu'on va passer d'une représentation à l'autre.
Pour cela, on construit de nouvelles variables Cj qui vont permettre de **maximiser la variance**.

Il existe des contraintes de simplicité : on parle de **combinaisons linéaires des variables initiales**.

$$C1 = A_1^1 X_1 + A_2^1 X_2 + \ldots + A_p^1 X_p$$

Géométriquement, ça donne :



Si on considère la nouvelle variable C, l'information est reconstituée de la manière la plus fiable possible au sens de la variance.

- **La première composante principale C1** se définit par la **combinaison linéaire des variables initiales** maximisant la variance.
- **La deuxième composante principale** : **maximise la variance**, et est **non-corrélée à la première composante** (principe de **l'orthogonalité**).
- Et ainsi de suite....

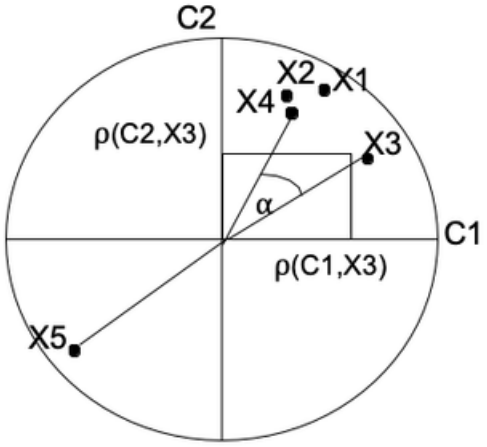
Au plus, on obtient **p composantes principales**.

En réalité, s'il existe une liaison entre les variables, **l'essentiel de l'information** (=la variance) est contenu dans les **premières composantes principales** *(en général, dans les 2 ou 3 premières composantes principales).*

L'analyse des liaisons entre les variables permet d'obtenir une **matrice de corrélation**.

Avec **p variables**, on obtient $\frac{p(p+1)}{2}$ **combinaisons possibles**.

Les liaisons se font 2 à 2, il n'y a **pas de liaisons multivariées**.



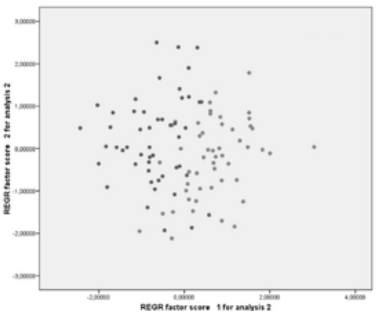
En ACP, on va représenter les variables sous la forme d'un **cercle des corrélations** *(C1 et C2 étant les deux premières composantes principales).*

On peut alors montrer que si des variables sont proches de la circonférence, alors **le cosinus de l'angle α est proche du coefficient ρ de corrélation entre ces 2 variables**.

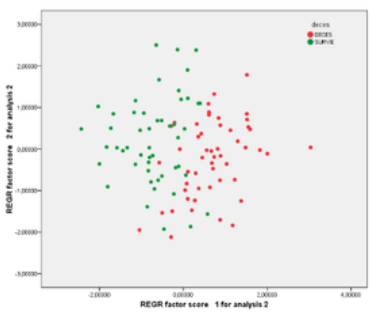
Exemple : infarctus du myocarde

- **Variables numériques** : fréquence cardiaque, index cardiaque, index systolique, pression diastolique, pression artérielle pulmonaire, pression ventriculaire, résistance pulmonaire
- **Variable qualitative** : décès *(pas prise en compte dans l'ACP en elle-même)*

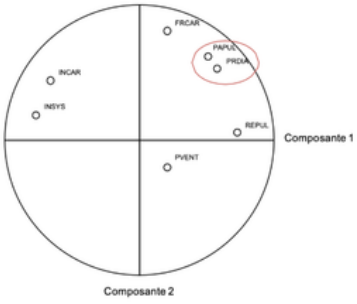
Ici, les objectifs vont être de vérifier la cohérence des données, rechercher les individus exceptionnels (en multivarié), rechercher l'existence de profils d'individus différents (sur p variables, donc en multivarié), et utiliser la variable « décès » comme variable illustrative.



Nuage d'individus



Nuage d'individus avec l'ajout d'une variable illustrative (vers l'inférentiel)



Cercle des corrélations entre les variables

STRATÉGIE D'ANALYSE

Statistiques descriptives	• Moyennes, pourcentages, intervalles de confiances, médianes • Graphiques (boxplot, histogramme)
Analyses univariées	Descriptives : statistiques et graphiques par groupes, survie (Kaplan- Meier)
	Tests statistiques (+- séries appariées) : • Pourcentages : Chi2, Fisher exact test • Moyennes : Student, ANOVA, Wilcoxon, Kruskal-Wallis • Corrélation de Pearson ou de Spearman • Log Rank (survie) • Interactions en fonction de la biologie • Séries chronologiques, corrélations spatiales…
Analyse multivariée	• Choix de la méthode (R linéaire, R logistique, Modèle de Cox…)

	• Choix des variables initiales
	  ◦ Variables connues dans la littérature
	  ◦ Variables avec un sens biologique
	  ◦ Variables p<0,2 ou p<0,25 pour les tests univariés

	• Méthode pas à pas, avec les interactions, choix du critère statistique

	• Garder
	  ◦ Les variables sélectionnées par la méthode pas à pas
	  ◦ Les variables biologiquement pertinentes

	• Vérification de la qualité du modèle • Interprétation du modèle final

FIN

Courage à vous tous, je crois fort en vous :)