Modèle Linéaire "Généralisé"

Julien Ah-Pine (julien.ah-pine@univ-lyon2.fr)

Université Lyon 2

M1 Informatique 2015-2016

Motivations

- Les modèles linéaires forment une famille de méthodes statistiques des plus utilisée pour modéliser et prédire des phénomènes (naturels, démographiques, sociologiques ...). Ainsi il fait partie des outils fondamentaux pour tout statisticien, analyste de données, chargé d'études, "data scientist" ...
- Il s'agit d'une famille de méthodes vaste qui permet d'analyser beaucoup d'ensembles de données qu'il soit homogène ou hétérogène (mélange de variables continues et discrètes).
- Dans la perspective de l'étude de techniques de fouille de données ou d'apprentissage automatique, ils consituent des méthodes classiques qu'il faut maîtriser avant d'aller vers d'autres approches.

Objectifs

- Rappeler des concepts en statistiques et probabilités permettant d'avoir une bonne compréhension des fondements des modèles linéaires.
- Les modèles linéaires généralisés forment une famille de méthode très vaste. Nous nous focaliserons essentiellement sur les méthodes dites de **régression** où il s'agit de prédire une variable continue.
- Nous verrons le cas classique (MV, MCO) ainsi que la plupart des outils statistiques permettant d'analyser les résultats de l'estimation d'un modèle linéaire (propriétés des estimateurs, tests d'hypothèses).
- Nous verrons également des cas plus spécifiques où il faut apporter des modifications à l'approche classique afin d'obtenir des modélisations plus efficaces (MCG, PCR, PLS).

Qu'est ce qu'un modèle linéaire

- Il s'agit de modéliser et de prédire une variable Y à partir d'une ou plusieurs variables $X^1, \ldots, X^p$.
- Y est la **variable à expliquer** (ou endogène ou dépendante ou cible).
- $X^1, \ldots, X^p$ sont les **variables explicatives** (ou exogènes ou indépendantes) ou covariables.
- On définit au préalable un **modèle** qui spécifie le **type de relation** que l'on suppose entre Y d'une part et les $X^1, \ldots, X^p$ d'autre part.
- Il existe plusieurs façon de modéliser la dépendence entre la variable endogène et les variables exogènes. Si le modèle est un polynôme de degré 1 en fonction des paramètres, on parle alors de **modèle linéaire**.

Les cas et extensions abordés dans ce cours

- On considère essentillement $Y \in \mathbb{R}$ (modèle de régression).
- Il existe dans ce cas plusieurs types de variables exogènes qui aboutissent à plusieurs types de modèles linéaires :
  - Régression sur variables continues : $X^1, \ldots, X^p$ sont toutes quantitatives **(régression linéaire)**.
  - Régression sur variables qualitatives :
    - ⋆ Si $X^1, \ldots, X^p$ sont toutes qualitatives on parle d'**analyse de la variance (ANOVA)**.
    - ⋆ Si $X^1, \ldots, X^p$ sont un mélange de variables quantitatives et qualitatives on parle d'**analyse de la covariance (ANCOVA)**
- ▷ Nous étudierons essentiellement la régression sur variables continues.
- ▷ A partir de n observations des variables à l'étude, nous verrons comment ajuster (ou estimer) les paramètres du modèle.
- ▷ Nous verrons si les hypothèses sous-jacentes aux modèles sont respectées par l'ajustement obtenu ou pas.
- ▷ Si ce n'est pas le cas, nous verrons des extensions de ces méthodes permettant de tenir compte de ces situations.

Organisation du cours

- 10 séances CM+TD de 1h45
- 4 séances de TP sur R de 1h45

- 1 CC (Dossier et programmation R par groupe de 2)
- 1 ET de 2h (écrit individuel)

- Les supports de cours disponibles sur le BV. Pour y accéder il faudra rejoindre le groupe `M1_INFO_ML-1516`.

Références

- Bases en probabilités et statistiques :
  - Lejeune, M., 2010. *Statistique, la Théorie et ses Applications*, Springer
  - Bertrand, F. et Maumy-Bertrand, M., 2011. *Statistique pour les scientifiques*, Dunod
  - Saporta, G., 2006. *Probabilites, Analyse des Données et Statistique*, Technip
- Plus centrées sur les modèles linéaires :
  - Cornillon, P-A. et Matzner-Lober, E., 2007. *Régression Théorie et Applications*, Springer
  - Cornillon, P-A. et Matzner-Lober, E., 2011. *Régression avec R*, Springer
  - Guyon, X., 2001. *Statistique et économétrie*, Ellipses
  - Seber, G.A.F et Lee, A.J. 2003. *Linear Regression Analysis*, Wiley

Rappel du Sommaire

A quoi peut faire référence le terme “statistique(s)” ?

1. Peut désigner tout ensemble de **données**, généralement numérique, associé à un ensemble d'individus ou d'objets. Exemple : les statistiques du chômage.
2. Peut désigner tout ensemble de **méthodes (ou de modèles)** permettant d'analyser ces ensembles de données. Exemple : les modèles linéaires font partie de la statistique.
3. Peut également désigner une **fonction mathématique** visant à estimer le paramètre d'un modèle statistique . Exemple : la moyenne empirique.

▷ Dans le cadre de ce cours, nous sommes concernées par ces trois points.

L'utilisation des probabilités en statistique

- Les données à notre disposition sont des observations relatives à la réalisation d'un phénomène présentant un caractère aléatoire.
- Pour tenir compte de cet incertitude, nous avons recours aux **probabilités** qui jouent donc un rôle fondamental en statistique.
- Un modèle statistique cherche à appréhender un phénomène aléatoire par le biais d'une fonction mathématique. Mais, cette fonction mathématique (polynôme de degré 1 dans le cadre de ce cours) est une simplification de ce phénomène. Ainsi “**tous les modèles sont faux ; certains sont utiles**” (G. Box).
- Les données observées représentent notre matière première à partir de laquelle nous cherchons donc à estimer les paramètres d'un modèle afin de reproduire au mieux le phénomène.

Ensemble fondamental et évènements

- Une **expérience aléatoire** consiste à réaliser une action permettant d'obtenir un résultat qui constitue une observation du phénomène étudié. Exemple : lancé d'un dé à 6 faces.
- Ω désigne l'**ensemble fondamental** (ou univers) qui est l'ensemble des valeurs possibles que peut prendre les résultats de l'expérience aléatoire. Exemple : $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Un **évènement** E est un sous-ensemble de Ω. $E \subset \Omega$. Dans ce cas, on ne s'intéresse qu'à un certain type de résultats de l'expérience aléatoire. Exemple : $E = \{1, 3, 5\}$ (le chiffre est impair).

Ensemble des évènements

- Dénotons par $\mathbb{E}$ l'ensemble des évènements[1] E construits à partir de Ω (ie l'ensemble des sous-ensembles de Ω).
- Soit $E, F \in \mathbb{E}$ deux évènement, on a sur $\mathbb{E}$ les opérations classiques de la théorie des ensembles :
  - La complémentarité : $\overline{E} = \Omega - E$ (dénoté également par $\Omega \setminus E$).
  - L'intersection : $E \cap F$.
  - L'union : $E \cup F$.
  - L'inclusion : $E \subseteq F$ si tous les éléments de E sont dans F.
- On dit que deux évènements E et F sont incompatibles si : $E \cap F = \emptyset$.

1. On supposera que $\mathbb{E}$ forme une tribu de Ω.

Mesure de probabilité

- On appelle **mesure de probabilité** une fonction $P : \mathbb{E} \to [0, 1]$ (ie une fonction qui associe à chaque E un nombre compris entre 0 et 1) vérifiant les axiomes de Kolmogorov :
  - $P(\Omega) = 1$
  - Pour tout ensemble dénombrable d'évènements incompatibles $E_1, E_2, \ldots, E_n$:

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

- On appelle le triplet $(\Omega, \mathbb{E}, P)$ un **espace probabilisé**.

Propriétés d'une mesure de probabilité

Pour tout $E, F \in \mathbb{E}$ on a :

- $P(E) \in [0, 1]$.
- $P(\overline{E}) = 1 - P(E)$.
- $P(E \cup F) = P(E) + P(F)$ si E et F sont incompatibles ie $E \cap F = \emptyset$.
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ dans le cas général.
- $(E \subseteq F) \Rightarrow (P(E) \leq P(F))$.

Par conséquent :

- $P(\Omega) = 1 \Rightarrow P(\emptyset) = 0$
  ($E = \Omega$ est l'évènement "certain" et $E = \emptyset$ est l'évènement "impossible").

Probabilité conditionnelle et indépendance entre évènements

- On définit la **probabilité conditionnelle** de E sachant F par :

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \text{ si } P(F) \neq 0$$

- On dit que deux évènements E et F sont **indépendants** ce qu'on dénote par $E \perp F$ ssi :

$$P(E \cap F) = P(E)P(F)$$

- Remarque : Si $E \perp F$ alors on a de façon équivalente la relation $P(E|F) = P(E)$. Autrement dit, en cas d'indépendance, la réalisation ou non de F n'a pas d'impact sur la réalisation ou non de E.
- On définit l'**indépendance mutuelle** entre un sous-ensemble d'évènements $E_1, \ldots, E_n$ de la façon suivante :

$$\forall I \subseteq \{1, \ldots, n\} : P(\bigcap_{i \in I} E_i) = \prod_{i \in I} P(E_i)$$

Variables aléatoires réelles (v.a.r.)

- Une **variable aléatoire** (v.a.) est une application $X : \Omega \to \mathbb{R}$ associée à une expérience aléatoire. Elle est définie sur l'espace fondamental et prend ses valeurs dans $\mathbb{R}$.
- Nous supposons qu'une v.a. est une **fonction mesurable** si pour tout intervalle I de $\mathbb{R}$, l'image réciproque $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\}$ existe et correspond à un évènement E de $\mathbb{E}$ [2].
- On distinque 2 types de v.a. :
  - Si X a des valeurs dans un ensemble dénombrable (ensemble des entiers typiquement) on parle de v.a. **discrète**.
  - Si X a des valeurs dans un ensemble non dénombrable (un intervalle de $\mathbb{R}$ typiquement) on parle de v.a. **continue**.

2. Dans le cas continu, cela implique que l'intervalle I appartient à la tribu borélienne.

Exemples de variables aléatoires

- Cas discret :
  - L'expérience aléatoire consiste à lancer 2 dés à 6 faces. $\Omega = \{1,2,3,4,5,6\}^2 = \{(1,1),(1,2),\ldots,(6,6)\}$. Si $X =$"Nombre d'apparition d'un 6" on a $X \in \{0,1,2\}$. Si $Y =$"Somme des chiffres" on a $Y \in \{2,\ldots,12\}$. Un exemple d'évènement est le sous-ensemble des résultats tels que la somme fasse 4 et dans ce cas $E = \{(1,3),(3,1),(2,2)\}$. On remarquera que E est équivalent à $Y = 4$.
- Cas continu :
  - L'expérience aléatoire consiste à mesurer la taille des patients d'un docteur. Dans ce cas, $\Omega =$"Ensemble des patients", $X =$"Taille en cm" et $X \in \mathbb{R}_+$.
  - L'expérience aléatoire consiste à prendre au hasard 10 étudiants et à déterminer leur moyenne. Ici, $\Omega =$"Ensemble des sous-ensembles de 10 étudiants", $X =$"Moyenne" et $X \in [0,20]$. Un exemple d'évènement est le sous-ensemble des éléments de Ω tels que la moyenne est supérieure à 12. On peut formalisé cet évènement par $E = \{X \geq 12\}$.

Mesure de probabilité et fonction de répartition d'une v.a.

- Soit X une v.a., on lui associe une mesure de probabilité P_X et qui pour tout intervalle I inclus dans $\mathbb{R}$ est définie par :

$$P_X(I) = P(X^{-1}(I))$$

où P est la mesure de probabilité sur $(\Omega, \mathbb{E})$.

- On dénote par F_X la **fonction de répartition** de X qui est définie par, $\forall x \in \mathbb{R}$:

$$\begin{aligned} F_X(x) &= P_X(]-\infty, x]) \\ &= P(X \leq x) \\ &= P(X^{-1}(]-\infty, x])) \end{aligned}$$

- La fonction de répartition permet de spécifier la loi de probabilité d'une v.a. X.

Propriétés d'une fonction de répartition

- F_X est **non décroissante** sur $\mathbb{R}$.
- F_X varie entre 0 et 1 quand x varie entre $-\infty$ et $+\infty$. En particulier :
  - $F_X(-\infty) = 0$
  - $F_X(+\infty) = 1$
- F_X est **continue à droite** et a une **limite à gauche** en tout $x \in \mathbb{R}$.
- On a les propriétés suivantes :
  - Pour tout $]a,b] \subset \mathbb{R}$ on a : $P_X(]a,b]) = P(a < X \leq b) = F_X(b) - F_X(a)$.
  - Pour tout $[a,b] \subset \mathbb{R}$ on a : $P_X([a,b]) = P(a \leq X \leq b) = F_X(b) - F_X(a) + P(X = a)$.
  - Pour tout $[a,b[\subset \mathbb{R}$ on a : $P_X([a,b]) = P(a \leq X < b) = F_X(b) - F_X(a) - P(X = b) + P(X = a)$.

V.a. discrète et fonctions de masse

- Si X est une v.a. discrète :
  - L'ensemble des valeurs possibles est dénombrable et peut être ordonné : $x_1 < x_2 \ldots < x_n$
  - F_X reste constante entre deux valeurs consécutives x_{i-1} et x_i, et présente un saut discontinu dès qu'elle atteint la valeur x_i.
  - En x_i, le saut observé est égal à la mesure de probabilité associée à ce point. Dans le cas discret, les mesures sont donc **concentrées** en les points définissant l'ensemble des valeurs possibles.
  - Pour tout x_i, on peut ainsi calculer sa probabilité $P_X(]x_{i-1}, x_i]) = p_X(x_i)$ que l'on appelle aussi **fonction de masse de probabilité** :

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1})$$

  - Nous avons donc :

$$F_X(x_i) = \sum_{j:x_j \leq x_i} p_X(x_j)$$

Quelques lois de probabilités discrètes

- La **loi de Bernouilli** correspond à observer le résultat d'une expérience aléatoire qui peut aboutir à deux états qui sont soit 0 ("échec") soit 1 ("succés"). On la dénote $\mathcal{B}(1,p)$ où $p \in [0,1]$ est la probabilité d'observer 1. Si $X \sim \mathcal{B}(1,p)$, on a :

$$p_X(1) = p \text{ et } p_X(0) = 1-p$$

- La **loi binomiale** $\mathcal{B}(n,p)$ consiste à réaliser n expériences de Bernouilli qui sont mutuellement indépendantes et à compter le nombre de succès. Si $X \sim \mathcal{B}(n,p)$, on a $\forall 0 \leq k \leq n$:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

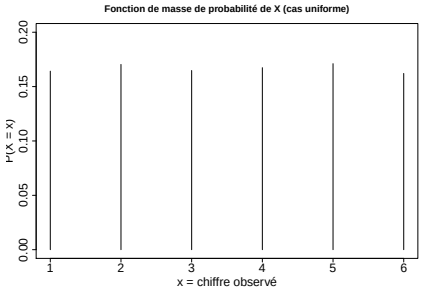
Quelques lois de probabilités discrètes (suite)

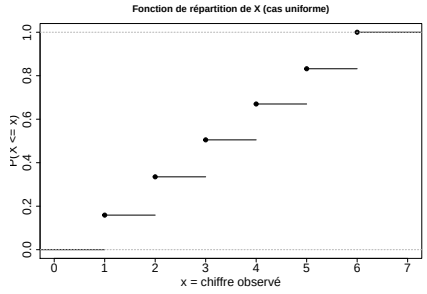
- La **loi multinomiale** $\mathcal{M}(n,\mathbf{p})$ est une généralisation de la loi binomiale dans le cas où l'expérience n'a pas deux mais $m > 2$ états distincts. Dans ce cas le paramètre $\mathbf{p} = (p_1, \ldots, p_m)$ est un vecteur d'ordre q qui somme à 1 et dont les composantes sont les probabilités d'observer les différents états. Si $X \sim \mathcal{M}(n,\mathbf{p})$, on a $\forall \mathbf{n} = (n_1, \ldots, n_m), \sum_j n_j = n$:

$$p_X(n_1, \ldots, n_m) = \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m}$$

Exemple empirique de fonctions de masse et de répartition

- Cas du lancé de dé qui suit une loi multinomiale (6 faces sont possibles) avec des probabilités uniformes (chaque face a une probabilité de 1/6), $\mathcal{M}(n, (\frac{1}{6} \quad \ldots \quad \frac{1}{6}))$.
- Simulation de n =10000 lancés et estimation empirique des fonctions de masse et de répartition.





V.a. continue et fonction de densité

- Si X est une v.a. continue :
  - Un point x n'a pas en soi de mesure de probabilité car il est vu tel un intervalle de longueur nulle. Ainsi, dans le cas continu $P_X(]x,x]) = 0$ pour tout x de $\mathbb{R}$.
  - En revanche, on définit une **fonction de densité de probabilité** notée f_X définie sur $\mathbb{R}$ et à valeur dans $\mathbb{R}_+$. $f_X(x)$ peut-être interprétée comme étant la mesure de probabilité d'appartenir à l'intervalle de longueur infinitésimale $]x, x+dt]$: $f_X(x) = P_X(]x, x+dt])$.
  - La fonction de répartition est alors définie par :

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

  - Nous avons également les définitions suivantes :

$$P_X(]a,b]) = \int_a^b f_X(t)dt, \quad \text{pour tout } a < b$$

$$P_X(I) = \int_{t \in I} f_X(t)dt, \quad \text{pour tout } I \subseteq \mathbb{R}$$

Quelques lois de probabilités continues

- La **loi uniforme** $\mathcal{U}(a, b)$ avec $a, b \in \mathbb{R}$ correspond à l'expérience aléatoire consistant à choisir au hasard un nombre compris entre a et b sachant qu'il y a équiprobabilité entre chaque nombre de cet intervalle. Si $X \sim \mathcal{U}(a, b)$, on a :

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

Quelques lois de probabilités continues (suite)

- La **loi normale ou de Laplace-Gauss** $\mathcal{N}(\mu, \sigma)$ avec $\mu, \sigma \in \mathbb{R}$. C'est une loi fondamentale à plusieurs titres. Elle caractérise des expériences aléatoires dont les résultats se répartissent de façon symétrique autour d'une moyenne en décrivant une forme ressemblant à une cloche (courbe de Gauss). $X \sim \mathcal{N}(\mu, \sigma)$ est telle que :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

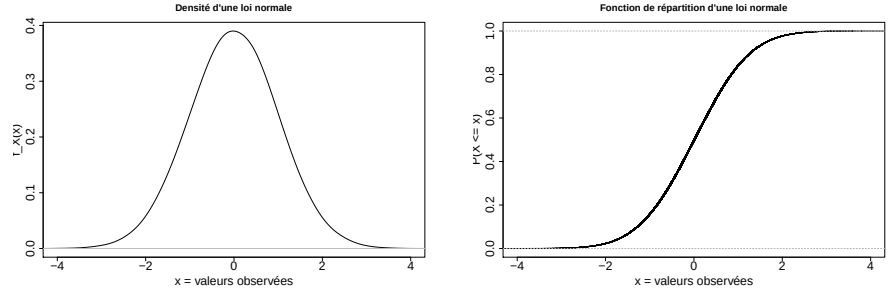
- On appelle **loi normale centrée réduite** le cas particulier $\mathcal{N}(0, 1)$. Dans ce cas la fonction de répartition est souvent notée ϕ et elle s'exprime comme suit :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t}{2}\right) dt$$

- En revanche, il n'y a pas d'expression analytique en termes de fonctions usuelles de la fonction de répartition d'une loi normale.

Exemple de fonctions de densité et de répartition

- Cas d'une loi normale centrée réduite $\mathcal{N}(0, 1)$.
- Simulation de 100000 observations et estimation empirique des fonctions de densité et de répartition (estimateur à noyau).



Quantile de niveau (ou d'ordre) α

- La notion de quantile d'une v.a. X est associée à la fonction de répartition de cette dernière.
- Le **quantile de niveau** $\alpha \in [0, 1]$, noté $q(\alpha)$, est le réel qui satisfait à la relation suivante :

$$P(X \leq q(\alpha)) = \alpha$$

ce qui est équivalent à

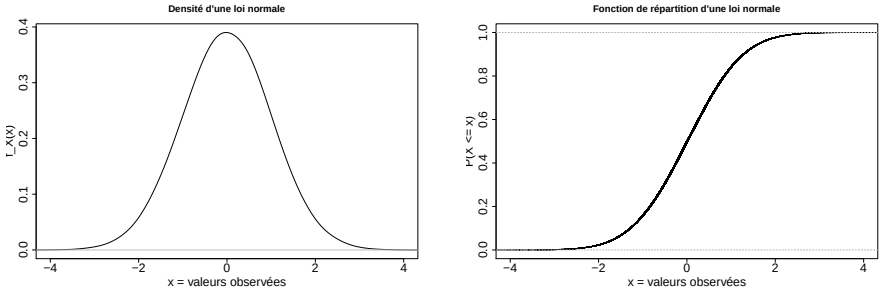
$$F_X(q(\alpha)) = \alpha$$

- C'est donc le réel tel que la probabilité que X lui soit inférieur vaut α.
- La médiane correspond par exemple au quantile de niveau 0.5.
- Dans le cas d'une v.a. discrète, F_X est discontinue et en forme d'escalier. Dans ce cas, on définit $q(\alpha)$ comme l'interpolation entre x_i et x_{i+1} tels que :

$$P(X \leq x_i) < \alpha \text{ et } P(X \leq x_{i+1}) > \alpha$$

Exemple de quantile

- Déterminez graphiquement une approximation de $q(0.8)$ pour $X \sim \mathcal{N}(0,1)$?



- Que peut-on dire à propos de $q(\alpha)$ et $q(1-\alpha)$ pour une loi de densité symmétrique par rapport à 0 ?

Fonction d'une variable aléatoire

- Soit X une v.a. de loi connue et soit Z une v.a. qui est fonction de X tel que $Z = g(X)$.
- Comment déterminer la loi de Z à partir de celle de X ?
- Dans le cas continu, on passe par la fonction de répartition. On exprime les évènements $Z \leq z$ en fonction de la variable X.
- Exemples :
  - Si $Z = 2X + 1$, on a :

$$\begin{aligned} P(Z \leq z) &= P(2X+1 \leq z) \\ &= P(X \leq \frac{z-1}{2}) \end{aligned}$$

    Donc $F_Z(z) = F_X((z-1)/2)$.
  - Si $Z = X^2$, on a :

$$\begin{aligned} P(Z \leq z) &= P(X^2 \leq z) \\ &= P(-\sqrt{z} \leq X \leq \sqrt{z}) \end{aligned}$$

    Donc $\forall z > 0 : F_Z(z) = F_X(\sqrt{z}) - F_X(-\sqrt{z})$.

Fonction d'une variable aléatoire (suite)

- Dans le cas continu, nous avons de façon générale :
  - Si $Z = g(X)$ avec g strictement croissante alors :
$$F_Z(z) = F_X(g^{-1}(z))$$
  - Si $Z = g(X)$ avec g strictement décroissante alors :
$$F_Z(z) = 1 - F_X(g^{-1}(z))$$
  - La densité de Z s'obtient ensuite par dérivation de F_Z.
- Dans le cas discret :
  - On utilise la fonction de masse car la fonction de répartition est en escalier ce qui rend son étude plus difficile.
  - L'ensemble des valeurs possibles pour Z sont $\{z_1, z_2, \ldots\}$ qui sont déterminées par $\{g(x_1), g(x_2), \ldots\}$.
  - On obtient $P(Z = z_k)$ en sommant les masses $P(X = g(x_i))$ telles que $g(x_i) = z_k$.

Espérance mathématique d'une v.a.

- L'**espérance mathématique** de la v.a. X est, si elle existe, la valeur réelle dénotée $\mathrm{E}(X)$ définie par :
  - si X est discrète :
$$\mathrm{E}(X) = \sum_i x_i p_X(x_i)$$
  - si X est continue :
$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$
- L'espérance sera aussi appelée **moyenne** et notée μ. Elle permet de mesurer une **tendance centrale** de la v.a. et s'exprime dans la même unité que cette dernière.

Espérance mathématique d'une fonction d'une v.a.

- Pour déterminer l'espérance de la v.a. $Z = g(X)$ à partir de la loi de X on peut procéder de deux façons :
  - Soit on détermine F_Z puis f_Z (p_Z dans le cas discret) comme vu précédemment et on utilise la loi de Z pour calculer $\mathrm{E}(Z)$ avec la formule précédente.
  - Soit on détermine $\mathrm{E}(Z)$ directement à partir de f_X (p_X dans le cas discret) à l'aide de la formule suivante :
    - ⋆ Dans le cas continu :

$$\mathrm{E}(Z) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$$

    - ⋆ Dans le cas discret :

$$\mathrm{E}(Z) = \sum_i g(x_i) p_X(x_i)$$

Linéarité de l'opérateur espérance mathématique

- Soit X une v.a. de loi connue et Z une v.a. définie comme étant une combinaison linéaire de fonctions de X :

$$Z = ag(X) + bh(X)$$

où a et b sont des réels et g et h des fonctions quelconques.

▷ Pour déterminer l'espérance de la v.a. Z à partir de la loi de X on peut appliquer la propriété de **linéarité** suivante :

$$\mathrm{E}(ag(X) + bh(X)) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(X))$$

- Rappel d'algèbre linéaire : une application g sur un espace vectoriel $\mathbb{F}$ est linéaire ssi, $\forall a, b \in \mathbb{R}$ et $\forall \mathbf{u}, \mathbf{v} \in \mathbb{F}$:

$$g(a\mathbf{u} + b\mathbf{v}) = ag(\mathbf{u}) + bg(\mathbf{v})$$

Variance mathématique d'une v.a.

▷ La **variance mathématique** de la v.a. X est, si elle existe, la valeur réelle dénotée $\mathrm{V}(X)$ définie par :

$$\begin{aligned} \mathrm{V}(X) &= \mathrm{E}((X - \mathrm{E}(X))^2) \\ &= \mathrm{E}((X - \mu)^2) \end{aligned}$$

- Remarque : $\mathrm{E}((X - \mu)^r)$ s'appelle le **moment centré d'ordre** r.

▷ L'**écart-type** de la v.a. X est la racine carrée de sa variance. Elle est notée σ et est formellement définie par :

$$\sigma(X) = \sqrt{\mathrm{V}(X)}$$

- $\mathrm{V}(X)$ et $\sigma(X)$ permettent de mesurer la dispersion autour de la moyenne de X. L'écart-type s'exprime dans la même unité que la v.a..
- Remarque : on utilisera parfois la notation σ_X.

Propriétés de la variance

▷ La variance peut également se calculer de la façon suivante :

$$\begin{aligned} \mathrm{V}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\ &= \mathrm{E}(X^2) - \mu^2 \end{aligned}$$

- Remarque : $\mathrm{E}(X^r)$ est appelé **moment d'ordre** r.

▷ Soit $Z = aX + b$. Pour calculer la variance de Z, on peut appliquer la propriété suivante :

$$\mathrm{V}(aX + b) = a^2 \mathrm{V}(X)$$

- Par ailleurs : $\mathrm{V}(X) = 0 \Leftrightarrow X$ est une v.a. certaine (probabilité 1 sur un seul point).

Couples de v.a.

- On suppose désormais deux v.a. X et Y et nous présentons des outils permettant de caractériser les relations entre ces deux variables. Ces deux v.a. forment un couple (X, Y) et leurs réalisations prennent valeurs dans $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ ie l'ensemble des couples de réels.
- Ainsi, la **mesure de probabilité** dite **jointe** $P_{X,Y}$ est une fonction portant sur les parties de $\mathbb{R}^2$.
- Dans la suite on ne considèrera que les couples de même type ie (discret, discret) ou (continue, continue).
- On appelle **fonction de répartition jointe** du couple (X, Y) que l'on note par $F_{X,Y}$, la fonction définie sur $\mathbb{R}^2$ par :

$$F_{X,Y}(x, y) = P((X \leq x) \cap (Y \leq y)) = P(X \leq x, Y \leq y)$$

- On retrouve la **fonction de répartition (marginale)** de chaque v.a. de la façon suivante :

$$F_X(x) = F_{X,Y}(x, +\infty) \text{ et } F_Y(y) = F_{X,Y}(+\infty, y)$$

Couple de v.a. discrètes

- Dans le cas où X et Y sont discrètes. On définit également :
  - La **fonction de masse de probabilités jointes** notée $p_{X,Y}$ définie par tout couple (x_i, y_j) par :

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$$

  - La **fonction de masse de probabilité marginale** de chaque v.a. que l'on peut déterminer à partir de $p_{X,Y}$ par :

$$\forall x_i : p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j) \text{ et } \forall y_j : p_Y(y_j) = \sum_i p_{X,Y}(x_i, y_j)$$

  - La **fonction de masse conditionnelle de X sachant Y**, $\forall y_j$:

$$\forall x_i : p_{X|Y}(x_i|y_j) = \frac{P_{X,Y}(X = x_i, Y = y_j)}{P_Y(Y = y_j)}$$

  - On peut de la même façon déterminer $p_{Y|X}$.

Couple de v.a. continues

- Dans le cas où X et Y sont continues, on a :
  - La **fonction de densité de probabilités jointes** notée $f_{X,Y}$ définie par tout couple $(x, y) \in \mathbb{R}^2$ à partir de la relation suivante :

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) ds dt$$

  - La **fonction de densité de probabilité marginale** de chaque v.a. est déterminée par :

$$\forall x \in \mathbb{R} : f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy$$

$$\forall y \in \mathbb{R} : f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dx$$

  - La **fonction de densité conditionnelle de X sachant Y**, $\forall y \in \mathbb{R}$:

$$\forall x \in \mathbb{R} : f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

  - On peut de la même façon déterminer $f_{Y|X}$.

Indépendance entre deux v.a.

▷ X et Y sont **indépendantes** ssi :

$$\forall x, y \in \mathbb{R}^2 : F_{X,Y}(x, y) = F_X(x) F_Y(y)$$

- Dans le cas discret, $X \perp Y$ se traduit aussi par :

$$\forall (x_i, y_j) : p_{X,Y}(x_i, y_j) = p_X(x_i) p_Y(y_j)$$

- Dans le cas continu, $X \perp Y$ se traduit plutôt par :

$$\forall (x, y) \in \mathbb{R}^2 : f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

- Si X et Y sont indépendantes, alors $g(X)$ et $h(Y)$ sont également indépendantes pour toutes fonctions (mesurables) g et h.

Espérance et covariance

- Nous pouvons généraliser l'espérance méthématique de $g(X)$ au cas d'un couple de v.a.. Soit $g(X, Y)$ une v.a. à valeurs dans $\mathbb{R}$:
  - Si X et Y sont discrètes :
$$\mathrm{E}(g(X,Y)) = \sum_i \sum_j g(x_i, y_j) p_{X,Y}(x_i, y_j)$$
  - Si X et Y sont continues :
$$\mathrm{E}(g(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) dxdy$$
- La propriété de **linéarité** de l'opérateur E se généralise également :
$$\mathrm{E}(ag(X) + bh(Y)) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(Y))$$
- On introduit un nouvel opérateur, la **covariance**, défini comme suit :
$$\mathrm{C}(X,Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)))$$
- La covariance permet d'appréhender la relation entre X et Y.

Covariance et indépendance

- On peut facilement montrer que :
$$\mathrm{C}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$
- On voit également facilement la propriété de symétrie de C :
$$\mathrm{C}(X,Y) = \mathrm{C}(Y,X)$$
- Nous avons ensuite les propriétés suivantes :
$$\begin{aligned} \mathrm{C}(aX + bY, Z) &= a\mathrm{C}(X,Z) + b\mathrm{C}(Y,Z) \\ \mathrm{C}(aX + b, cY + d) &= ac\mathrm{C}(X,Y) \end{aligned}$$
- En cas d'indépendance, nous avons la propriété remarquable suivante :
$$X \perp Y \Rightarrow \mathrm{C}(X,Y) = 0$$

  Attention ! la réciproque est fausse. En revanche nous dirons que X et Y sont **non corrélés** si $\mathrm{C}(X,Y) = 0$.

Variance et indépendance

- On remarquera le lien suivant entre la covariance et la variance :
$$\mathrm{C}(X,X) = \mathrm{V}(X)$$
- Ensuite, nous avons la propriété :
$$\mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{C}(X,Y)$$
- En cas d'indépendance, nous avons donc :
$$X \perp Y \Rightarrow \mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y)$$

Coefficient de corrélation linéaire

- Une mesure statistique très utilisée pour apprécier la relation entre deux variables X et Y est le **coefficient de corrélation linéaire** noté $\rho(X,Y)$ et défini par :
$$\rho(X,Y) = \frac{\mathrm{C}(X,Y)}{\sigma(X)\sigma(Y)}$$
- Nous avons les propriétés suivantes :
  - $\rho(X,Y) = \rho(Y,X)$
  - $\rho(aX + b, cY + d) = \rho(X,Y)$
  - $X \perp Y \Rightarrow \rho(X,Y) = 0$
  - **Attention !** ici aussi, $\rho(X,Y) = 0 \not\Rightarrow X \perp Y$
- Le coefficient de corrélation est une mesure bornée quelque soit les lois de X et Y, ce qui facilite son interprétation :
$$-1 \leq \rho(X,Y) \leq 1$$
- La coefficient est dit linéaire car il atteint les bornes -1 ou 1 lorsqu'il existe une dépendance liénaire entre X et Y ie du type $Y = aX + b$.

n-uplet de v.a. ou vecteurs aléatoires

- On considère maintenant un nombre quelconque de v.a. $X_1, X_2, \ldots, X_n$ ($n = 2$ correspondait au cas précédent).
- Nous parlerons alors de **vecteurs aléatoires** étant donné que l'on concatène ces v.a. en une entité X :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

- ▷ Formellement un vecteur aléatoire X est une application de l'espace probabilisé $(\Omega, \mathbb{E}, P)$ dans l'espace vectoriel $\mathbb{R}^n$ (muni de sa tribu borélienne), $X : \Omega \to \mathbb{R}^n$.
- ▷ X représente un vecteur colonne de taille $(n \times 1)$ mais nous écrirons également de façon équivalente $X = (X_1, \ldots, X_n)$.
- ▷ **Attention !** aux notations, X peut désigner soit une v.a. soit un vecteur aléatoire. Dans le dernier cas, $\mathbf{x}$ désignera une réalisation de X. Ainsi $\mathbf{x} = (x_1, \ldots, x_n)$ est un vecteur de n réels.

n-uplet de v.a. ou vecteurs aléatoires

- Nous avons les extensions des définitions suivantes pour un vecteur aléatoire X, $\forall \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$:
  - **Fonction de répartition jointe** :

$$F_X(\mathbf{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

  - **Fonction de densité jointe** :

$$f_X(\mathbf{x}) = \frac{\partial^n F}{\partial x_1 \ldots \partial x_n}(x_1, \ldots, x_n)$$

  - Chaque compostante X_i de X est une v.a.. Dans ce cas, sa **fonction de densité (marginale)** [3] est définie par :

$$\forall x \in \mathbb{R} : f_{X_i}(x) = \int_{\mathbb{R}^{n-1}} f_X(x_1, \ldots, x_i, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

3. Fonction de masse marginale dans le cas discret.

Vecteur espérance d'un vecteur aléatoire

- ▷ Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire tel que chacune de ses composantes admette une espérance. On appelle **vecteur espérance** de X le vecteur noté $\mathrm{E}(X)$ et défini par :

$$\mathrm{E}(X) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}$$

- Si on dénote $\mu_i = \mathrm{E}(X_i)$, alors on écrira également :

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

- ▷ Propriété de **linéarité** de E dans le cas multidimensionnel. Soit $\mathbf{A}$ une matrice carrée de taille $(n \times n)$ et $\mathbf{b}$ un vecteur de taille $(n \times 1)$ alors :

$$\mathrm{E}(\mathbf{A}X + \mathbf{b}) = \mathbf{A}\mathrm{E}(X) + \mathbf{b}$$

Matrice de variance-covariance d'un vecteur aléatoire

- ▷ Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire. Si chaque couple (X_i, X_j) admet une covariance alors on appelle **matrice de variance-covariance** de X, la matrice carrée d'ordre n, notée $\mathrm{V}(X)$ et définie par :

$$\mathrm{V}(X) = \mathbf{\Sigma}_X = \begin{pmatrix} \mathrm{V}(X_1) & \mathrm{C}(X_1, X_2) & \ldots & \mathrm{C}(X_1, X_n) \\ \mathrm{C}(X_2, X_1) & \mathrm{V}(X_2) & \ldots & \mathrm{C}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{C}(X_n, X_1) & \mathrm{C}(X_n, X_2) & \ldots & \mathrm{V}(X_n) \end{pmatrix}$$

- ▷ Propriétés de la matrice de variance-covariance :
  - $\mathbf{\Sigma}_X$ est symmétrique réelle
  - $\mathbf{\Sigma}_X = \mathrm{E}(XX^\top) - \mu\mu^\top$ où $X^\top$ est le vecteur transposé de X.
  - Soit $\mathbf{A}$ une matrice carrée de taille $(n \times n)$ et $\mathbf{b}$ un vecteur de taille $(n \times 1)$ alors :

$$\mathrm{V}(\mathbf{A}X + \mathbf{b}) = \mathbf{A}\mathrm{V}(X)\mathbf{A}^\top$$

Vecteurs aléatoires gaussiens (ou loi normale multivariée)

- $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire gaussien si toute combinaison linéaire de ses composantes suit une loi normale univariée.
  - **Attention !** Il ne suffit pas que chaque composante suive une loi normale pour que $\mathbf{x}$ soit un vecteur gaussien. En revanche, si chaque composante suit une loi normale et si elles sont indépendantes deux à deux[4] alors X est un vecteur gaussien.
  - Un vecteur aléatoire gaussien X d'ordre n est défini par son vecteur espérance μ et sa matrice de variance-covariance $\mathbf{\Sigma}$ qui sont également tous deux d'ordre n.

4. ie tout couple (X_i, X_j) est tel que $X_i \perp X_j$.

Vecteurs aléatoires gaussiens (suite)

- Si $X \sim \mathcal{N}_n(\mu, \mathbf{\Sigma})$, on a la fonction de densité (multidimensionnelle) suivante, $\forall \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$:

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(det(\mathbf{\Sigma}))^{n/2}} \exp\left(-\frac{1}{2}\left((\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right)$$

  où $det(\mathbf{\Sigma})$ et $\mathbf{\Sigma}^{-1}$ sont le déterminant et l'inverse de $\mathbf{\Sigma}$.
- Notons que $\mathbf{\Sigma}$ doit être inversible (elle est donc de rang n) sinon le vecteur gaussien n'appartiendrait pas à $\mathbb{R}^n$.
- Par ailleurs nous avons les propriétés suivantes concernant $\mathbf{\Sigma}$:
  - $\mathbf{\Sigma} = \mathbf{\Sigma}^\top$ (symmétrie).
  - $\mathbf{\Sigma}$ est définie positive ie que $\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0} : \mathbf{x}^\top \mathbf{\Sigma} \mathbf{x} > 0$.

Propriétés des vecteurs aléatoires gaussiens

- Dans le cas de vecteurs gaussiens nous avons la propriété que la covariance nulle entre deux composante (X_i, X_j) implique l'indépendance entre ces deux composantes. Autrement dit si $X \sim \mathcal{N}_n(\mu, \mathbf{\Sigma})$ alors non corrélation est synonyme d'indépendance. **Attention !** Nous avons déjà précisé que ceci est faux dans le cas général.
- Ainsi, les composantes d'un vecteur gaussien sont indépendantes ssi $\mathbf{\Sigma}$ est diagonale.
- Soit $X \sim \mathcal{N}_n(\mu, \mathbf{\Sigma})$ de dimension n et soit le vecteur aléatoire $Z = \mathbf{A}X + \mathbf{b}$ où $\mathbf{A}$ une matrice de taille $(m \times n)$ de rang m avec $m \leq n$ et $\mathbf{b}$ un vecteur de taille $(m \times 1)$ alors :

$$Z \sim \mathcal{N}_m(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$$

Echantillonage et vecteur aléatoire i.i.d.

- On suppose que l'on étudie un phénomène aléatoire dont les valeurs possibles sont dans Ω et qui suit une loi de probabilité P **inconnue**.
- On s'intéresse plus particulièrement à une caractéristique de ce phénomène qui est représentée par une v.a. X de loi P_X qui est donc également **inconnue**.
- Notre but est de déterminer des propriétés de X (et donc de Ω). On cherche alors à appréhender la loi P_X.
- Pour ce faire, l'idée est de répéter des expériences aléatoires afin d'obtenir des observations de X et à partir desquelles on va inférerer des propriétés. C'est à ce niveau que l'on rentre dans la discipline des statistiques en comparaison de celle des probabilités.
- Formellement, on suppose que l'on réalise n expériences aléatoires qui sont **mutuellement indépendantes**. Autrement dit, les conditions ou les résultats d'un sous-ensemble expériences ne doivent pas impacter celles d'un autre sous-ensemble d'expériences.

Echantillonage et vecteur aléatoire i.i.d. (suite)

- On associe à chaque expérience aléatoire $i = 1, \ldots, n$, une v.a. X_i qui est une réplique de la v.a. X. Les X_i suivent donc la même loi que X appelée **loi mère**.
- Les v.a. $X_1, \ldots, X_n$ sont donc **indépendantes** (les expériences aléatoires associées à chacune d'entre elles étant indépendantes) et **identiquement distribuées à la v.a. parente** X. On écrira que les v.a. sont **i.i.d.**.
- On appelle un **échantillon aléatoire** de taille n un vecteur aléatoire $(X_1, \ldots, X_n)$ où les X_i sont des v.a. i.i.d. de variable parente X.

▷ **Attention !** Il faut faire la distinction entre :

- $(X_1, \ldots, X_n)$ qui est un vecteur aléatoire de v.a. i.i.d. donc un échantillon et,
- $(x_1, \ldots, x_n)$ qui est un vecteur de réels réprésentant les observations (ie les réalisations des v.a.).

Exemples d'echantillons

- Le phénomène étudié est la pollution de l'air d'une ville.
  - Ω représente par exemple tous les états possibles de la pollution de l'air.
  - On s'intéresse en particulier à la concentration en ozone dans l'air. On représente cette caractéristique du phénomène par la v.a. X.
  - Pour étudier X, on décide de relever de façon indépendante n mesures de la concentration en ozone.
  - Les relevés doivent être mutuellement indépendants. Chaque mesure x_i est la réalisation d'une v.a. X_i qui suit la loi de X.
- Le phénomène est le chômage en France à un instant donné.
  - Ω représente tous les français au chômage à l'instant choisi.
  - On s'intéresse à la durée de chômage. On représente cette caractéristique par une v.a. X.
  - N'ayant pas les données sur toute la population on décide d'effectuer un sondage sur n individus.
  - On tire au hasard et de façon indépendante n individus indexés par $i = 1, \ldots, n$. Au $i$ème individu, ω_i, tiré au hasard on lui associe une v.a. $X_i =$ "durée du chômage".

Statistique de l'échantillon et modèle statistique

- Soit $(X_1, \ldots, X_n)$ un échantillon de variable parente X.

▷ On appelle **statistique** (de l'échantillon) toute v.a. T qui est une fonction des v.a. i.i.d. $X_1, \ldots, X_n$.

▷ Un **modèle statistique** de X est une description formelle cherchant à approximer X. Le modèle statistique est défini par la donnée de deux éléments :

- L'ensemble des valeurs prise par X.
- Un ensemble de lois de probabilités $\mathbb{H}$ auquel peut appartenir P_X.

- On distingue trois types de modèles statistiques :
  - ▷ Les modèles **paramétriques** : lorsque $\mathbb{H}$ est une famille de probabilités (comme la loi normale) qui dépend d'un ensemble fini de paramètres $\mathbb{P}$ dont le domaine Θ est un sous-ensemble de $\mathbb{R}^p$.
  - Les modèles **non paramétriques** : lorsque les éléments de $\mathbb{H}$ ne prennent pas de forme pré-déterminée. C'est le cas si on considère Θ de dimension infinie.
  - Les modèles **semi paramétriques** : lorsque les éléments de $\mathbb{H}$ ont une composante paramétrique et une composante non paramétrique.

Moyenne, variance et moments empiriques

▷ On appelle **moyenne de l'échantillon** $(X_1, \ldots, X_n)$ ou **moyenne empirique**, la statistique notée $\overline{X}$ et définie par :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

▷ La **variance empirique** est la statistique notée S^2 et définie par :

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Le **moment empirique d'ordre** r et le **moment centré empirique d'ordre** r sont les statistiques notées M_r et MC_r et définies par :

$$M_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r \text{ et } MC_r = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^r$$

- Les moments empiriques permettent de définir des estimateurs des propriétés de X.

Convergence d'une suite de v.a.

- On considère une suite infinie de v.a. $\{X_1, \ldots, X_n, \ldots\}$ qui sont toutes définies sur le même espace probabilisé. On note de façon brève cette suite par $\{X_n\}_{n\geq 1}$.

▷ Remaque : dans le cas général, les v.a. d'une suite ne sont pas forcément i.i.d. donc $\{X_n\}_{n\geq 1}$ n'est pas nécessairement un échantillon.

- Notons par F_{X_n} la fonction de répartition de X_n.
- On peut définir plusieurs modes de convergence pour une suite $\{X_n\}_{n\geq 1}$ vers une v.a. X défini sur le même espace probabilisé.
- On dit que $\{X_n\}_{n\geq 1}$ **converge en loi** vers la v.a. X si en tout x où la fonction de répartition F_X est continue on a la propriété suivante :

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

- On dit que $\{X_n\}_{n\geq 1}$ **converge en probabilité** vers la v.a. X si :

$$\forall \epsilon > 0 : \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

Convergence d'une suite de v.a. (suite)

- La convergence en probabilité implique la convergence en loi.
- La convergence en loi est souvent utilisée en pratique car elle permet d'approximer la fonction de répartition de X quand l'échantillon est de très grande taille.

▷ Remarque : de façon générale, il est claire que plus nous avons d'observations du phénomène étudié (grands échantillons), meilleure est la modélisation statistique de ce dernier.

- Les **statistiques asymptotiques** est la branche de la discipline qui consiste à étudier les propriétés des outils statistiques lorsque $n \to \infty$.

Loi des grands nombres

- Il existe des résultats théoriques forts en statistiques asymptotiques connus sous le terme de théorèmes limites.

Théorème. (La loi faible des grands nombres)

Soit $\{X_n\}_{n\geq 1}$ une suite de v.a. de même loi sur un même espace probabilisé, deux à deux indépendantes, ayant une espérance μ et une variance σ^2. Notons par $\overline{X_n}$ l'espérance empirique de $\{X_n\}_{n\geq 1}$ défini par :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Alors la suite $\{\overline{X_n}\}_{n\geq 1}$ converge en probabilité vers la v.a. constante et égale à μ.

Théorème de limite centrale

Théorème. (Théorème de limite centrale)

Soit $\{X_n\}_{n\geq 1}$ une suite de v.a. i.i.d., ayant une espérance μ et une variance σ^2. Notons :

$$Y_n = \frac{\overline{X_n} - \mu}{\sqrt{\sigma^2/n}}$$

Alors la suite $\{Y_n\}_{n\geq 1}$ converge en loi vers une v.a. de loi $\mathcal{N}(0,1)$.

Théorème. (Théorème de limite centrale dans le cas multivarié)

Soit $\{X_n\}_{n\geq 1}$ une suite de vecteurs aléatoires de dimension p i.i.d., ayant un vecteur espérance μ et une matrice de variance-covariance $\mathbf{\Sigma}$. Notons :

$$Y_n = \sqrt{n}(\overline{X_n} - \mu)$$

Alors la suite $\{Y_n\}_{n\geq 1}$ converge en loi vers un vecteur aléatoire de loi $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$.

Rappel du Sommaire

Notations

- Y est une v.a. réelle qui représente la variable à expliquer ie la caractéristique du phénomène qui nous intéresse.
- $X^1, \ldots, X^p$ sont des v.a. réelles qui représentent les variables explicatives ie des caractéristiques autres (du même phénomène ou d'autres phénomènes) que l'on suppose avoir un impact sur Y.
- $X = (X^1, \ldots, X^p)$ est un vecteur aléatoire de taille $(p \times 1)$.
- $P(X)$ est la fonction de probabilité de X.
- $P(Y|X)$ est la probabilité conditionnelle de Y sachant X.
- $\mathrm{E}(X^j)$ est l'espérance de la variable X^j.
- $\mathrm{E}_X(f(X))$ est l'espérance de $f(X)$ par rapport à X.
- $\mathrm{E}_{Y|X}(f(Y)|X)$ est l'espérance de $f(Y)$ par rapport à Y sachant X.
- On sintéresse au couple (X, Y). On suppose un échantillon de taille n où chaque élément est indicé par $i = 1, \ldots, n$. L'échantillon s'écrit donc $((X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n))$. Ainsi les (X_i, Y_i) sont i.i.d. selon $P(X, Y)$ la loi jointe du couple.

Notations (suite)

- La réalisation de ces n expériences élatoires conduisent à instancier une table des données $\mathbf{X}$ de taille $(n \times p)$ et un vecteur colonne $\mathbf{y}$ de taille $(n \times 1)$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{1p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- La ligne i de $\mathbf{X}$ est associée au vecteur aléatoire X_i qu'on appelle également l'observation ou l'individu i.
- La colonne j de $\mathbf{X}$ est associée à la v.a. X^j qu'on appelle également la variable ou l'attribut j. L'ensemble des variables exogènes sera noté $\mathbb{A}$.
- $x_{ij}, y_i \in \mathbb{R}$ sont respectivement les termes généraux de $\mathbf{X}$ et de $\mathbf{y}$ qui sont les valeurs de la variable explicative j et celle de la variable à expliquer, prises par l'individu i.

Notations (suite)

- Chaque observation i est associée à un vecteur colonne $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ de taille $(p \times 1)$ appartenant à $\mathbb{R}^p$ (correspondant à la ligne i de $\mathbf{X}$).
- Chaque variable explicative j est associée à un vecteur colonne $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})$ de taille $(n \times 1)$ des valeurs observées pour la variable j (correspondant à la colonne j de $\mathbf{X}$).
- $\mathbf{y} = (y_1, \ldots, y_n)$ est le vecteur colonne de taille $(n \times 1)$ des valeurs observées pour la variable à expliquer Y.
- Nous dénoterons par X le vecteur aléatoire associé à un individu quelconque, $\mathbf{x}$ son vecteur des valeurs des variables exogènes dans $\mathbb{R}^p$ et y sa valeur de la variable endogène dans $\mathbb{R}$.

Exemple introductif : Ozone

- Exemple tiré de [Cornillon et al (2011)].
- Le phénomène : la pollution de l'air, l'influence de l'ozone sur la santé.
- Le statisticien : prévision des pics de concentration de l'ozone (O_3).
- Modèle : prévision de O_3 en fonction de la température (T_{12}), du vent (VX) et de la nébulosité (NE_{12}).
- Extrait des données observées :

Observation	O_3	T_{12}	VX	NE_{12}
1	63.6	13.4	9.35	7
2	89.6	15	5.4	4
3	79	7.9	19.3	8
4	81.2	13.1	12.6	7
5	88	14.1	-20.3	6
⋮	⋮	⋮	⋮	⋮

Modélisation linéaire

- L'objectif est d'estimer une fonction $f(X) \approx Y$ qui représente la relation entre Y et X.
- Dans l'exemple précédent $X = (T_{12}, VX, NE_{12})$ et $Y = O_3$.
- La modélisation **linéaire** consiste à prendre pour **hypothèse** que la relation f est un **polynôme de degré 1 de ses paramètres**. On suppose donc que :

$$f(X) = a_1 X^1 + \ldots + a_p X^p$$

- Dans ce cas, l'ensemble des paramètres est $\mathbb{P} = \{a_1, \ldots, a_p\}$.
- Si on note $\mathbf{a} = (a_1, \ldots, a_p)$ on peut écrire matriciellement :

$$f(X) = X^\top \mathbf{a}$$

où $X^\top$ est le vecteur transposé de $X = (X^1, \ldots, X^p)$ (c'est donc un vecteur ligne de taille $(1 \times p)$).

- $\mathbf{a} \in \mathbb{R}^p$ est appelé **vecteur des paramètres** des coefficients du modèle.

Modélisation linéaire (suite)

- On remarque que $f(\mathbf{0}) = 0$ ce qui implique que f décrit un hyperplan passant forcément par l'origine.
- On préfèrera souvent à ce type de modèle un modèle dit **affine** possédant une ordonnée à l'origine qui soit distincte de 0.
- On supposera donc par la suite que notre modèle s'écrit comme suit :

$$f(X) = a_0 + a_1 X^1 + \ldots + a_p X^p$$

- Cela revient à ajouter une **variable explicative** X^0 qui est en fait la **constante** 1. On a $\mathbb{A} = \{X^0, X^1, \ldots, X^p\}$. Le vecteur $\mathbf{a}$ devient un élement de $\mathbb{R}^{p+1}$. De même, la matrice $\mathbf{X}$ des données observées est désormais de taille $(n \times (p+1))$:

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} \quad ; \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{1p} \\ \vdots & \vdots & \ldots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

Modélisation linéaire (suite)

- En résumé, nous supposons dans la suite que $\forall i = 1, \ldots, n$:

$$f(X_i) = X_i^\top \mathbf{a}$$

où $X = (X^0, \ldots, X^p)$ est un vecteur aléatoire et $\mathbf{a} = (a_0, \ldots, a_p)$ un vecteur de paramètres tous deux de $\mathbb{R}^{p+1}$.

- **Attention !** C'est en fonction des paramètres que la fonction doit être un polynôme de degré 1 pour que le modèle soit linéaire. Ainsi le modèle suivant dit modèle additif généralisé est également linéaire :

$$f(X) = a_1 f_1(X^1, \ldots, X^p) + \ldots + a_m f_m(X^1, \ldots, X^p)$$

où $f_1, \ldots, f_m$ sont des fonctions usuelles comme par exemple $f_j(X^1, \ldots, X^p) = \prod_{k=1}^{j} X^k$ ou $f_j(X^1, \ldots, X^p) = \log(X^j)$.

- Il existe des extentions du modèle linéaire qui consiste à faire des **expansions de base**.

Tout modèle est faux !

- Tout modèle est une tentative de formalisation de la réalité. Tout modèle fait donc des erreurs. Il en va de même pour le modèle linéaire.
- Il faut donc poser :

$$Y = \underbrace{a_0 + a_1 X^1 + \ldots + a_p X^p}_{f(X)} + \epsilon$$

où ϵ est appelé **résidu ou erreur**.

- Autrement dit, pour que l'égalité soit vraie, il faut intégrer une **variable résiduelle** qui mesure la différence entre la réalité Y et le modèle $f(X)$.

Modèle linéaire gaussien et hypothèses sur ϵ

- Sous l'hypothèse de linéarité, nous avons $\forall i = 1, \ldots, n$:

$$Y_i = X_i^\top \mathbf{a} + \epsilon_i$$

- Nous supposerons que chaque ϵ_i est une v.a. représentant le résidu associé à l'individu i.
- Nous faisons à présent les **hypothèses probabilistes** suivantes concernant les résidus :
  - **Moyenne nulle**, $\forall i : \mathrm{E}(\epsilon_i) = 0$.
  - **Homoscédasticité**, $\forall i : \mathrm{V}(\epsilon_i) = \sigma^2$.
  - **Normalité**, $\forall i : \epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
  - **Non corrélation**, $\forall i \neq j : \mathrm{C}(\epsilon_i, \epsilon_j) = 0$.
- Soit le vecteur des résidus $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Les hypothèses précédentes, dénotées par $\mathcal{H}_\epsilon$ s'écrivent matriciellement :

$$\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

où $\mathbf{0}$ est le vecteur $(n \times 1)$ rempli de 0 et $\mathbf{I}_n$ est la matrice identité d'ordre n.

Modèle linéaire gaussien et hypothèses sur ϵ (suite)

- $\mathcal{H}_\epsilon$ revient à supposer que $(\epsilon_1, \ldots, \epsilon_n)$ est un vecteur gaussien.
- On déduit également de $\mathcal{H}_\epsilon$, les propriétés importantes suivantes, $\forall i = 1, \ldots, n$:
  - $Y_i | X_i \sim \mathcal{N}(X_i^\top \mathbf{a}, \sigma^2)$
  - $\mathrm{E}_{Y_i|X_i}(Y_i|X_i) = X_i^\top \mathbf{a}$
  - $\mathrm{V}(Y_i) = \sigma^2$
  - $\mathrm{C}(Y_i, Y_j) = 0$
- Autrement dit, pour le couple (X, Y) à l'étude nous supposons fondamentalement que :

$$\mathrm{E}_{Y|X}(Y|X) = X^\top \mathbf{a}$$

- Par ailleurs, l'ensemble complet des paramètres du modèle gaussien devient : $\mathbb{P} = \{\mathbf{a}, \sigma\}$ où $\mathbf{a} \in \mathbb{R}^{p+1}$ et $\sigma \in \mathbb{R}_+$.

Maximisation de la vraisemblance

- Dans le cadre d'un modèle statistique, une méthode d'inférence classique permettant d'estimer les paramètres d'un modèle est la **maximisation de la vraisemblance (MV)**.
- La **vraisemblance ("likelihood")** est la probabilité d'observer l'échantillon :

$$vr(\mathbf{a}, \sigma^2) = P(Y_1, \ldots, Y_n | X_1, \ldots, X_n; \mathbf{a}, \sigma^2)$$

- L'**estimateur du MV** est la valeur des paramètres qui maximise la probabilité d'observer l'échantillon. On résoud donc le problème :

$$\max_{(\mathbf{a}, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}} \prod_{i=1}^{n} P(Y_i | X_i; \mathbf{a}, \sigma^2)$$

Maximisation de la vraisemblance

- Rappelons que dans le cadre du modèle gaussien, nous avons fait l'hypothèse de normalité :

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_i | X_i \sim \mathcal{N}(X_i^\top \mathbf{a}, \sigma^2)$$

- Comme les Y_i sont également indépendantes, nous avons :

$$\begin{aligned} vr(\mathbf{a}, \sigma^2) &= P(Y_1, \ldots, Y_n | X_1, \ldots, X_n; \mathbf{a}, \sigma^2) \\ &= \prod_{i=1}^{n} P(Y_i | X_i; \mathbf{a}, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - X_i^\top \mathbf{a}}{\sigma}\right)^2\right) \end{aligned}$$

Vraisemblance et Log-vraisamblance (suite)

- Il est plus commode de maximiser, de manière équivalente, le **logarithme de la vraisemblance** :

$$lvr(\mathbf{a}, \sigma^2) = \log(\prod_{i=1}^{n} P(Y_i | X_i; \mathbf{a}, \sigma^2)) = \sum_{i=1}^{n} \log(P(Y_i | X_i; \mathbf{a}, \sigma^2))$$

- Dans le modèle gaussien cela se réduit à :

$$\begin{aligned} lvr(\mathbf{a}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - X_i^\top \mathbf{a}}{\sigma}\right)^2\right)\right) \\ &= \sum_{i=1}^{n} \left(-\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2}\left(\frac{Y_i - X_i^\top \mathbf{a}}{\sigma}\right)^2\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - X_i^\top \mathbf{a}\right)^2 \end{aligned}$$

Estimateurs du MV

- Rappelons que $\mathbb{P} = \{\mathbf{a}, \sigma\}$.
- L'estimateur du MV de $\mathbf{a}$ est défini de la façon suivante :

$$\hat{\mathbf{a}}_{mv} = \arg \max_{\mathbf{a} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \log(P(Y_i | X_i; \mathbf{a}, \sigma^2))$$

- Tandis que l'estimateur du MV de σ^2 est tel que :

$$\hat{\sigma}^2_{mv} = \arg \max_{\sigma^2 \in \mathbb{R}} \sum_{i=1}^{n} \log(P(Y_i | X_i; \hat{\mathbf{a}}_{mv}, \sigma^2))$$

- Il faut donc résoudre des problèmes d'optimisation afin d'obtenir les estimations du MV des paramètres du modèle.
- Dans le cas classique, nous obtenons une solution analytique comme nous allons le voir.

Rappels en calcul différentiel

- Si $f : \mathbb{R}^{p+1} \to \mathbb{R}$ est différentiable, alors la fonction ∇f défini par :

$$\nabla f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f}{\partial a_0}(\mathbf{a}) \\ \frac{\partial f}{\partial a_1}(\mathbf{a}) \\ \vdots \\ \frac{\partial f}{\partial a_p}(\mathbf{a}) \end{pmatrix}$$

  est appelé **gradient** de f.
- ∇f est une fonction de $\mathbb{R}^{p+1}$ dans $\mathbb{R}^{p+1}$ et peut être vue comme un **champ de vecteurs** (fonction qui associe à tout point un vecteur).
- Voici quelques formules de dérivations dans le cas multivarié. La dérivée est calculée par rapport à la variable $\mathbf{x}$. $\mathbf{A}$ est une matrice de réels de taille $(m \times n)$ et $\mathbf{y}$ un vecteur de réels de taille $(m \times 1)$:
  - Si $f(\mathbf{x}) = \mathbf{y}^\top \mathbf{A}\mathbf{x}$ ou si $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y}$ alors $\nabla f(\mathbf{x}) = \mathbf{A}^\top \mathbf{y}$
  - Si $\mathbf{A}$ est carrée et $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ alors $\nabla f(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$
  - Si $\mathbf{A}$ est carrée symétrique et $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ alors $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Rappels en optimisation non contrainte

Théorème. (Condition nécessaire du 1er ordre (CNPO) (point intérieur))

Soit $\mathbb{S}$ un sous-ensemble de $\mathbb{R}^{p+1}$ et f une fonction de classe $\mathcal{C}^1$ de $\mathbb{S}$ dans $\mathbb{R}$. Si $\mathbf{x}^*$ est un optimiseur local de f sur $\mathbb{S}$ et si $\mathbf{x}^*$ est un point intérieur alors on a :

$$\nabla f(\mathbf{x}^*) = \mathbf{0}$$

On dit alors que $\mathbf{x}^*$ est un **point critique** *(ou stationnaire)*

- Ainsi pour déterminer les estimateurs du MV, il faut commencer par déterminer les points critiques.
- Il faut ensuite vérifier qu'il s'agisse bien d'un maximiseur et pour cela il faut étudier le signe de la matrice hessienne (CNSO). On admettra dans la suite qu'il sagit bien d'un maximiseur.

Estimateur du MV de $\mathbf{a}$

- Dans le modèle gaussien, la log-vraisemblance appliquée aux données de l'échantillon $\mathbf{X}$ et $\mathbf{y}$ s'écrit :

$$lvr(\mathbf{a}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{a}\right)^2$$

- En utilisant les notations matricielles, nous avons :

$$lvr(\mathbf{a}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)^\top \left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)$$

- Pour déterminer l'estimateur du MV de $\mathbf{a}$, il faut calculer le gradient de *lvr* par rapport à $\mathbf{a}$.
- On voit alors que seul le troisième terme de droite de l'équation précédente dépend de $\mathbf{a}$.
- Posons $c = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$, la *lvr* s'écrit alors :

$$lvr(\mathbf{a}, \sigma^2) = c - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)^\top \left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)$$

Estimateur du MV de $\mathbf{a}$ (suite)

- On développe la fonction *lvr* et on obtient :

$$\begin{aligned} lvr(\mathbf{a}, \sigma^2) &= c - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)^\top \left(\mathbf{y} - \mathbf{X}\mathbf{a}\right) \\ &= c - \frac{1}{2\sigma^2}\left(\left(\mathbf{y}^\top - (\mathbf{X}\mathbf{a})^\top\right)\left(\mathbf{y} - \mathbf{X}\mathbf{a}\right)\right) \\ &= c - \frac{1}{2\sigma^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{a} - (\mathbf{X}\mathbf{a})^\top\mathbf{y} + (\mathbf{X}\mathbf{a})^\top\mathbf{X}\mathbf{a}\right) \\ &= c - \frac{1}{2\sigma^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{a} - \mathbf{a}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{a}^\top\mathbf{X}^\top\mathbf{X}\mathbf{a}\right) \end{aligned}$$

- En appliquant les règles de dérivation matricielle, on obtient :

$$\begin{aligned} \nabla_{\mathbf{a}} lvr = \mathbf{0} &\Leftrightarrow 2\mathbf{X}^\top\mathbf{X}\mathbf{a} - 2\mathbf{X}^\top\mathbf{y} = \mathbf{0} \\ &\Leftrightarrow \hat{\mathbf{a}}_{mv} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \end{aligned}$$

- Remarque : nous supposons que $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ existe. Nous apportons des précisons sur cette hypothèse plus loin.

Estimateur du MV de σ

- On s'intéresse cette fois-ci au paramètre σ^2.
- La *lvr* est une fonction réelle de σ^2 :

$$lvr(\mathbf{a}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{a}\right)^2$$

- En dérivant *lvr* par rapport à σ^2, on obtient :

$$\begin{aligned} \frac{\partial lvr}{\partial \sigma^2} = 0 &\Leftrightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{a}\right)^2 = 0 \\ &\Leftrightarrow \hat{\sigma}^2_{mv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \hat{\mathbf{a}}_{mv})^2 \end{aligned}$$

Hypothèse sur la matrice $\mathbf{X}$

- Nous supposons qu'il y a plus d'observations que de variables et donc : $n > p+1$.
- Plus haut nous avons également supposé que la matrice $\mathbf{X}^\top\mathbf{X}$ était non singulière (**inversible**). On suppose donc que $\mathbf{X}$ est de **plein rang**. En d'autres termes $rg(\mathbf{X}) = p+1$, ou encore la base de l'espace vectoriel engendré par les colonnes de $\mathbf{X}$ est de dimension $p+1$.
- Le cas où $\mathbf{X}$ n'est pas de plein rang se présente lorsqu'il existe des variables exogènes qui sont colinéaires ou, de façon plus générale, lorsqu'il existe des dépendances linéaires entre ces variables.
- Si $\mathbf{X}^\top\mathbf{X}$ est singulière alors il existe une infinité de $\hat{\mathbf{a}}_{mv}$, les coefficients ne sont pas uniques : le problème n'est pas **identifiable**.
- L'hypothèse que $\mathbf{X}$ est de plein rang sera notée $\mathcal{H}_{\mathbf{X}}$.
- Ainsi, en outre de l'hypothèse de la dépendance linéaire entre X et Y, le modèle linéaire gaussien fait deux autres hypothèses fondamentales que sont $\mathcal{H}_\epsilon$ et $\mathcal{H}_{\mathbf{X}}$.

Prédiction du modèle gaussien

- Une fois estimé $\hat{\mathbf{a}}_{mv}$ on peut calculer les prédictions du modèle pour un quelconque $\mathbf{x} \in \mathbb{R}^{p+1}$:

$$\hat{f}(\mathbf{x}) = \mathbf{x}^\top\hat{\mathbf{a}}_{mv} = \mathbf{x}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

- Pour calculer l'erreur de prédiction on regarde ce que prédit le modèle estimé sur les observations données par les lignes de $\mathbf{X}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}}_{mv} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

où $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ est le **vecteur des valeurs prédites**.

- L'estimation de l'**erreur de prédiction** est donc donnée par :

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

où $\|.\|$ est la norme euclidienne.

Lien entre estimateur MV et estimateur MCO

- Rappelons l'expression de la *lvr* :

$$lvr(\mathbf{a}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top\mathbf{a}\right)^2$$

- La partie de la *lvr* dépendante de $\mathbf{a}$ est en rouge. Celle-ci est appelée **somme des carrés des résidus** :

$$scr(\mathbf{a}) = \sum_{i=1}^{n}\underbrace{(y_i - \overbrace{\mathbf{x}_i^\top\mathbf{a}}^{f(x_i)})^2}_{\epsilon_i^2} = \left(\mathbf{y} - \mathbf{X}^\top\mathbf{a}\right)^\top\left(\mathbf{y} - \mathbf{X}^\top\mathbf{a}\right)$$

- L'**estimateur des moindres carrés ordinaires (MCO)** est :

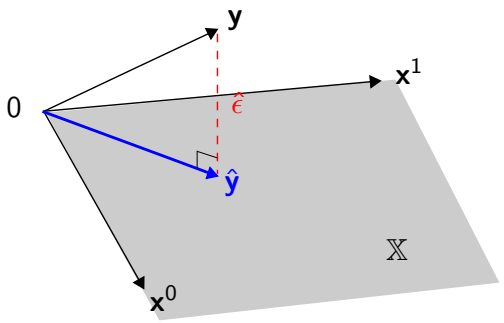
$$\hat{\mathbf{a}}_{mco} = \arg\min_{\mathbf{a}\in\mathbb{R}^{p+1}} scr(\mathbf{a})$$

- Clairement $\max lvr(\mathbf{a}) \Leftrightarrow \min scr(\mathbf{a})$. On a donc la propriété suivante :

$$\hat{\mathbf{a}}_{mco} = \hat{\mathbf{a}}_{mv}$$

Interprétations géométriques des MCO

- Interprétation géométrique :



$$\hat{\mathbf{y}} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top}_{\text{Opérateur de projection}}\mathbf{y} = \mathbf{P}_{\mathbb{X}}\mathbf{y}$$

où $\mathbf{P}_{\mathbb{X}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ est l'opérateur de projection sur $\mathbb{X} = Vect\{\mathbf{x}_0, \ldots, \mathbf{x}_p\}$

- Les MCO consistent à projeter orthogonalement $\mathbf{y}$ sur $\mathbb{X}$, le sous-espace vectoriel (sev) de $\mathbb{R}^{p+1}$ engendré par $\{\mathbf{x}^0, \ldots, \mathbf{x}^p\}$.
- Intuitivement, comme on cherche à minimiser $scr(\mathbf{a})$, on voit que la plus courte distance entre $\mathbf{y}$ et le sous-espace est donnée par la projection orthogonale.

Quel estimateur ? Rappels en statistique inférentielle

- Nous venons de voir qu'il existait potentiellement plusieurs méthodes pour définir des estimateurs des paramètres d'un modèle statistique.
- Comment choisir de façon générale un estimateur par rapport à un autre ?
- Nous faisons maintenant des rappels en statistique inférentielle permettant de donner des critères de comparaison entre estimateurs.
- Nous décrirons ces concepts du point de vue général et nous abandonnons pendant quelques slides le problème particulier du modèle linéaire gaussien.

Définition d'un estimateur

- Soit, de manière générale, $\mathcal{L}(\theta)$, une fonction de probabilité quelconque dépendant du paramètre θ.
- Soit $L_1, \ldots, L_n$, n v.a. i.i.d. de loi parente $\mathcal{L}(\theta)$ (ie un échantillon aléatoire).
- Un **estimateur (ponctuel)** de θ est une v.a. T qui est fonction des v.a. $\{L_i\}_{i=1}^n$ et qui permet d'estimer θ. Il s'agit donc d'une statistique de l'échantillon.
- Si on dispose des réalisations $l_1, \ldots, l_n$ des v.a., on peut alors déterminer une **estimation** (i.e. une valeur de l'etimateur) de θ. On écrit dans ce cas $\hat{\theta} = T(l_1, \ldots, l_n) = \hat{T}$.

Propriétés classiques d'un estimateur

- T est dit **convergent** s'il converge en probabilité vers θ :

$$\forall \epsilon > 0 : P(|T(L_1, \ldots, L_n) - \theta| > \epsilon) \to 0 \text{ lorsque } n \to +\infty$$

- T est dit **sans biais** si :

$$\forall n : \mathrm{E}(T(L_1, \ldots, L_n)) = \theta$$

- T est dit **efficace** si parmi les estimateurs convergents et sans biais, il est celui de variance minimale.

Fonction de risque quadratique

- Pour mesurer l'erreur entre l'estimateur T et le paramètre θ, une fonction usuelle est le **risque quadratique** défini par :

$$R(T, \theta) = \mathrm{E}_T((T - \theta)^2)$$

- Un estimateur T sera dit **meilleur** qu'un estimateur T' si $R(T, \theta) < R(T', \theta)$.
- Un estimateur T' est **inadmissible** s'il existe au moins un estimateur T qui lui est meilleur. Dans le cas contraire il est dit admissible.
- On a la décomposition suivante du risque quadratique :

$$R(T, \theta) = \mathrm{V}(T) + (\mathrm{E}(T) - \theta)^2$$

- Les estimateurs **sans biais** et les estimateurs **précis** (de variance faible) sont donc intéressants.
- Une stratégie classique en statistique consiste à chercher des estimateurs sans biais puis parmi cette sous-classe d'estimateurs, on cherche les estimateurs de variance minimale.

Espérance de $\mathbf{a}_{mv}$

- L'estimateur du MV de $\mathbf{a}$ est noté $\mathbf{a}_{mv}$:

$$\mathbf{a}_{mv} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

- $\mathbf{a}_{mv}$ est sans biais :

$$\begin{aligned} \mathrm{E}_{Y|X}(\mathbf{a}_{mv}|\mathbf{X}) &= \mathrm{E}_{Y|X}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y|\mathbf{X}\right) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathrm{E}_{Y|X}\left(Y|\mathbf{X}\right) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{a} \\ &= \mathbf{a} \end{aligned}$$

Variance de $\mathbf{a}_{mv}$

- L'estimateur du MV de $\mathbf{a}$ est noté $\mathbf{a}_{mv}$:

$$\mathbf{a}_{mv} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

- Variance de $\mathbf{a}_{mv}$:

$$\begin{aligned} \mathrm{V}_{Y|X}(\mathbf{a}_{mv}|\mathbf{X}) &= \mathrm{V}_{Y|X}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y|\mathbf{X}\right) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathrm{V}_{Y|X}\left(Y|\mathbf{X}\right)\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \end{aligned}$$

Théorème de Gauss-Markov

Théorème. (Théorème de Gauss-Markov)

Sous l'hypothèse que le vecteur des résidus $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ soit tel que $\mathrm{E}(\epsilon) = \mathbf{0}$ et $\mathrm{V}(\epsilon) = \sigma^2\mathbf{I}_n$, l'estimateur $\mathbf{a}_{mv} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y$ est, parmi les estimateurs **linéaires sans biais**, *celui de* **variance minimale**.

- Ce résultat montre l'**efficacité** des estimateurs du MV dans le cadre du modèle linéaire.

Espérance de σ^2_{mv}

- L'estimateur du MV de σ^2, noté σ^2_{mv}, est donné par :

$$\sigma^2_{mv} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^\top \mathbf{a}_{mv}\right)^2$$

- On montre que l'espérance de σ^2_{mv} vaut :

$$\mathrm{E}_{Y|X}(\sigma^2_{mv}) = \frac{(n-(p+1))}{n}\sigma^2$$

- L'estimateur du MV de la variance résiduelle est donc **biaisée** et une estimation non biaisée est alors :

$$\frac{1}{n-(p+1)}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\mathbf{a}}_{mv}\right)^2$$

Espérance de σ^2_{mco}

- Nous interprétons maintenant l'estimateur de σ^2 par les MCO.
- $\epsilon = Y - f(X)$ est le vecteur des résidus $(\epsilon_1, \ldots, \epsilon_n)$.
- Nous avons : $\epsilon = Y - \mathbf{P}_{\mathbb{X}} Y = (\mathbf{I}_n - \mathbf{P}_{\mathbb{X}})Y$. où $\mathbf{I}_n$ est la matrice d'identité et $\mathbf{P}_{\mathbb{X}}$ est la matrice de projection orthogonale sur $\mathbb{X}$ le sev engendré par les variables appartenant à $\mathbb{A}$.
- Par ailleurs : $\mathbf{P}_{\mathbb{X}^\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbb{X}}$ est l'opérateur de projection sur $\mathbb{X}^\perp$ le sev orthogonal à $\mathbb{X}$.
- On voit que : $\epsilon = \mathbf{P}_{\mathbb{X}^\perp}\epsilon$ (puisque $\epsilon \in \mathbb{X}^\perp$).
- Il est facile de montrer que $\mathrm{E}(\epsilon) = \mathbf{0}$ ($\mathbf{a}_{mco}$ étant sans biais).
- Nous remarquons ensuite que :

$$\begin{aligned}
\mathrm{E}(\|\epsilon\|^2) &= \mathrm{E}(\epsilon^\top \epsilon) \text{ (où } \|.\| \text{ est la norme euclidienne)} \\
&= \mathrm{E}((\mathbf{P}_{\mathbb{X}^\perp}\epsilon)^\top \mathbf{P}_{\mathbb{X}^\perp}\epsilon) \\
&= \mathrm{E}(\epsilon^\top \mathbf{P}_{\mathbb{X}^\perp}\epsilon) \text{ (car } \mathbf{P}^\top_{\mathbb{X}^\perp} = \mathbf{P}_{\mathbb{X}^\perp} \text{ et } \mathbf{P}^2_{\mathbb{X}^\perp} = \mathbf{P}_{\mathbb{X}^\perp})
\end{aligned}$$

Espérance de σ^2_{mco} (suite)

- ...

$$\begin{aligned}
\mathrm{E}(\|\epsilon\|^2) &= \mathrm{E}(\sum_{i,j=1}^{n} \epsilon_i \epsilon_j [\mathbf{P}_{\mathbb{X}^\perp}]_{ij}) \\
&= \sum_{i,j=1}^{n} \mathrm{E}(\epsilon_i \epsilon_j)[\mathbf{P}_{\mathbb{X}^\perp}]_{ij} \\
&\quad \text{(car E est linéaire et } \mathbf{P}_{\mathbb{X}^\perp} \text{ est connue } \mathbf{X} \text{ étant donnée)} \\
&= \sigma^2 \sum_{i=1}^{n} [\mathbf{P}_{\mathbb{X}^\perp}]_{ii} \text{ (car } \mathrm{E}(\epsilon_i \epsilon_i) = \sigma^2 \text{ et } \mathrm{E}(\epsilon_i \epsilon_j) = 0 \text{ si } i \neq j) \\
&= \sigma^2(n - (p+1)) \\
&\quad \text{(car } \sum_{i=1}^{n} [\mathbf{P}_{\mathbb{X}^\perp}]_{ii} = Tr(\mathbf{P}_{\mathbb{X}^\perp}) = dim(\mathbb{X}^\perp) = n - \underbrace{dim(\mathbb{X})}_{p+1})
\end{aligned}$$

Espérance de σ^2_{mco} (suite)

- ... Nous remarquons finalement que :

$$\mathrm{E}(\|\epsilon\|^2) = \sigma^2(n - (p+1))$$

- On en déduit l'estimation MCO $\hat{\sigma}^2_{mco}$ non biaisée suivante :

$$\begin{aligned}
\hat{\sigma}^2_{mco} &= \frac{\mathrm{E}(\|\hat{\epsilon}\|^2)}{n - (p+1)} \\
&= \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \hat{\mathbf{a}}_{mco})^2}{n - (p+1)}
\end{aligned}$$

Choix des estimateurs pour la suite et notations

- Compte tenu des résultats précédents, nous nous intéresserons dans la suite aux estimateurs des MCO.
- Pour alléger les notations, nous noterons :
  - $\mathbf{a}^*$ l'estimateur des coefficients de la régression est donc :

$$\hat{\mathbf{a}}^* = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

  - σ^{*2} l'estimateur non biaisé de la variance résiduelle est donc :

$$\hat{\sigma}^{*2} = \frac{\|\mathbf{y} - \mathbf{X}^\top \hat{\mathbf{a}}^*\|^2}{n - (p+1)}$$

- Nous cherchons maintenant à étendre les résultats précédents en déterminant les lois de probabilités de $\mathbf{a}^*$ et σ^{*2}.
- Pour cela nous faisons d'abord quelques rappels en probabilité concernant les lois dérivées de la loi normale.

Rappels sur les lois de probabilité

- La loi du χ^2_p (**Chi2**) à p degrés de liberté (d.d.l.) est la loi de la somme :

$$Y_1^2 + \ldots + Y_p^2 \text{ où les } Y_i \text{ sont i.i.d. selon une } \mathcal{N}(0,1)$$

- Si Y est un vecteur aléatoire de taille $(p \times 1)$ suivant une $\mathcal{N}_p(\mathbf{0}, \Sigma)$ et si Σ est inversible alors $Y^\top \Sigma^{-1} Y \sim \chi^2_p$.
- Soit $Y = (Y_1, \ldots, Y_n)$ un vecteur aléatoire tel que les Y_i sont des v.a. i.i.d. selon $\mathcal{N}(m, \sigma^2)$. Soit $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, la moyenne empirique des Y_i et $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$, la variance empirique de l'échantillon. Nous avons les résultats suivants :

$$\overline{Y} \sim \mathcal{N}(m, \tfrac{\sigma^2}{n})$$
$$S^2 \sim \chi^2_{n-1}$$
$$\overline{Y} \text{ et } S^2 \text{ sont indépendants}$$

Rappels sur les lois de probabilité (suite)

- La loi de $\mathcal{T}_p$ (**Student**) à p d.d.l. est la loi du quotient :

$$\frac{Y}{\sqrt{Z/p}} \text{ où } Y \sim \mathcal{N}(0,1) \text{ et } Z \sim \chi^2_p \text{ et } Y \text{ et } Z \text{ sont indépendants}$$

- La loi de $\mathcal{F}_{p,q}$ (**Fisher**) à p et q d.d.l. est la loi du quotient :

$$\frac{Z_1/p}{Z_2/q} \text{ où } Z_1 \sim \chi^2_p \text{ et } Z_2 \sim \chi^2_q \text{ et } Z_1 \text{ et } Z_2 \text{ sont indépendants}$$

- Le carré d'une loi de Student $\mathcal{T}_p$ est une loi de Fisher $\mathcal{F}_{1,p}$

Lois des estimateurs

- A l'aide des rappels précédents, nous pouvons établir les lois suivantes.
- Sous les hypothèses $\mathcal{H}_{\mathbf{X}}$ et $\mathcal{H}_\epsilon$ et en supposant σ^2 connue, on a :

$$\mathbf{a}^* \sim \mathcal{N}_{p+1}(\mathbf{a}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

$$(n-(p+1))\frac{\sigma^{*2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

$$\mathbf{a}^* \text{ et } \sigma^{*2} \text{ sont indépendants}$$

- Sous les hypothèses $\mathcal{H}_{\mathbf{X}}$ et $\mathcal{H}_\epsilon$ mais en supposant σ^2 inconnue :

$$\forall j = 1, \ldots, p : T_j = \frac{a_j^* - a_j}{\hat{\sigma}^* \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} \sim \mathcal{T}_{n-(p+1)} \text{ (où } \hat{\sigma}^* = \sqrt{\hat{\sigma}^{*2}})$$

Estimation ponctuelle *versus* estimation par intervalle

- Nous avons étudié jusqu'à présent des **estimations ponctuelles** ie **une unique valeur estimée** du paramètre.
- Compte tenu du fait que notre modèle est une approximation de la réalité, il est préférable de donner un **ensemble de valeurs** au sein duquel nous estimons (avec une certaine confiance) que le paramètre s'y trouve. On parle d'**intervalle de confiance (IC)** pour un paramètre.
- Dans la recherche d'un IC, l'estimation ponctuelle sert de base. C'est autour de cette valeur, que l'on va chercher un IC.
- Nous faisons dans ce qui suit des rappels concernant l'estimation paramétrique par IC.

Rappels sur les intervalles de confiance

- Soit, de manière générale, $\mathcal{L}(\theta)$, une fonction de probabilité quelconque dépendant du paramètre θ.
- Soit $L_1, \ldots, L_n$ un échantillon aléatoire de taille n de loi parente $\mathcal{L}(\theta)$.
- Soit T un estimateur ponctuel de θ dont on connaît la loi P pour chaque valeur de $\theta_0 \in \Theta$. On prendra bien sûr le meilleur estimateur que l'on ait pour θ.
- Etant donné $\theta_0 \in \Theta$, on appelle **intervalle de probabilité de niveau** $1-\alpha$ **de** T, tout couple de bornes (t_1, t_2), tel que :

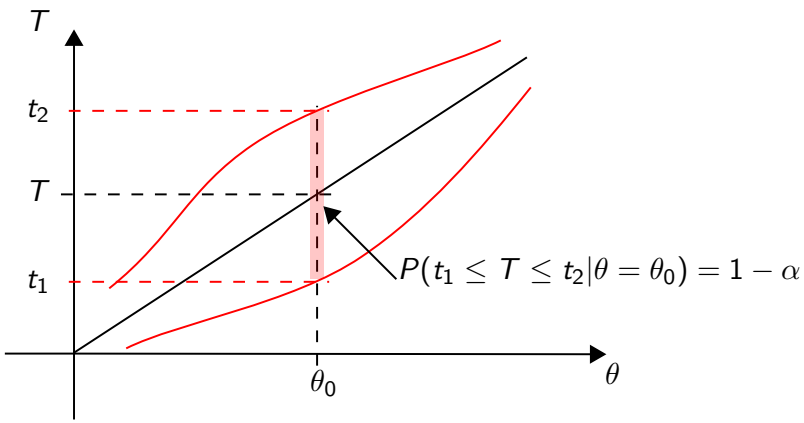
$$P(t_1 \leq T \leq t_2 | \theta = \theta_0) = 1 - \alpha$$

où $\alpha \in [0, 1]$ est le **risque** que T ne soit pas dans $[t_1, t_2]$.

- On remarquera que t_1 et t_2 dépendent de θ_0.
- On suppose que l'on peut définir pour **chaque valeur** $\theta_0 \in \Theta$ un tel intervalle.

Rappels sur les intervalles de confiance (suite)

- Intervalle de probabilité de niveau $1-\alpha$ issu de $P(T|\theta)$.



Rappels sur les intervalles de confiance (suite)

- Soient $l_1, \ldots, l_n$ des réalisations des v.a. $L_1, \ldots, L_n$.
- Ces réalisations permettent d'avoir à leur tour une réalisation $T(l_1, \ldots, l_n)$ (notée aussi $\hat{T}$) de T.
- On appelle alors **intervalle de confiance de niveau** $1-\alpha$ **de** θ, noté $\mathbb{IC}_{1-\alpha}(\theta)$, tout intervalle $[\theta_1, \theta_2]$ tel que :

$$P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha$$

- On notera également : $\mathbb{IC}_{1-\alpha}(\theta) = [\theta_1, \theta_2]$.
- θ_1 et θ_2 dépendent de l'estimation $T(l_1, \ldots, l_n)$ qui elle même dépend de la valeur de θ (précédemment on raisonnait avec $P(T|\theta)$).
- Ainsi $[\theta_1, \theta_2]$ est un intervalle aléatoire et nous voyons plus loin une méthode classique pour déterminer des IC.

Rappels sur les intervalles de confiance (suite)

- α représente la probabilité que θ ne soit pas dans l'IC.
- Cette probabilité peut être répartie de part et d'autre de θ_1 et θ_2.
- Posons $\alpha = \alpha_1 + \alpha_2$ où :
  - α_1 mesure le risque que la vraie valeur de θ soit **au-dessous de** θ_1.
  - α_2 mesure le risque que la vraie valeur de θ soit **au-dessus de** θ_2.
- On dit alors que l'IC est **bilatéral** si $\alpha_1 \neq 0$ **et** $\alpha_2 \neq 0$. Si de plus, $\alpha_1 = \alpha_2 = \alpha/2$ alors on dit que l'IC est **symmétrique**. Sinon il est dit **dissymétrique**.
- On dit au contraire que l'IC est **unilatéral** si $\alpha_1 = 0$ **ou (xor)** $\alpha_2 = 0$. Dans ce cas, on ne s'intéresse qu'à une seule borne :
  - Si $\alpha_1 = 0$ alors $\mathbb{IC}_{1-\alpha}(\theta)$ est de type $]-\infty, t_2]$ (borne maximale).
  - Si $\alpha_2 = 0$ alors $\mathbb{IC}_{1-\alpha}(\theta)$ est de type $[t_1, -\infty[$ (borne minimale).

Rappels sur les intervalles de confiance (suite)

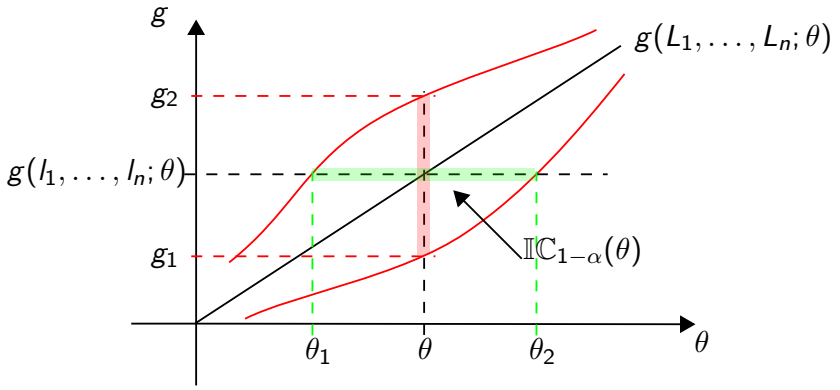
- La **méthode de la fonction pivot** permet de déterminer $[\theta_1, \theta_2]$ (IC de θ) à partir de $[t_1, t_2]$ (intervalle de probabilité de T).
- Une fonction $g(l_1, \ldots, l_n; \theta)$ est appelée **fonction pivot** si :
  - La loi de g est connue.
  - Pour tous réels g_1, g_2 tels que $g_1 \leq g_2$ et tout $(l_1, \ldots, l_n) \in \mathbb{R}^n$, la double inégalité $g_1 \leq g(l_1, \ldots, l_n; \theta) \leq g_2$ peut se résoudre (ou "pivoter") en θ afin d'obtenir un IC de ce dernier :

$$\theta_1(l_1, \ldots, l_n) \leq \theta \leq \theta_2(l_1, \ldots, l_n)$$

- L'existence d'une fonction pivot permet de déterminer des IC de niveau donné quelconque.
- Autrement, la méthode asymptotique faisant usage du TLC pourra être appliquée de façon approximative (car n est en pratique fini).

Rappels sur les intervalles de confiance (suite)

- Illustration de la détermination d'un IC de niveau $1-\alpha$ pour θ par "pivotage" de la fonction g.



- On a : $\theta_1 = g_2^{-1}(g(l_1, \ldots, l_n; \theta))$ et $\theta_2 = g_1^{-1}(g(l_1, \ldots, l_n; \theta))$.

Rappels sur les intervalles de confiance (suite)

- Exemple : $L_1, \ldots, L_n$ sont i.i.d. de loi mère $\mathcal{N}(\mu, 1)$.
- On souhaite déterminer un IC pour le paramètre μ.
- $T = \overline{L} = \frac{1}{N}\sum_i L_i$ est le meilleur estimateur de μ.
- On définit la fonction pivot suivante :

$$g(L_1, \ldots, L_n; \mu) = \frac{\overline{L} - \mu}{1/\sqrt{n}}$$

- On sait que $g(L_1, \ldots, L_n; \mu) \sim \mathcal{N}(0, 1)$ et donc :

$$P(t_1 \leq g(L_1, \ldots, L_n; \mu) \leq t_2) = 1 - \alpha$$

où $t_1 = u(\frac{\alpha}{2})$ et $t_2 = u(1 - \frac{\alpha}{2})$ où $u(\alpha)$ est le quantile de niveau α d'une v.a. $X \sim \mathcal{N}(0, 1)$ (ie $P(X \leq u(\alpha)) = \alpha$).

Rappels sur les intervalles de confiance (suite)

- Soit $\mathbf{l} = (l_1, \ldots, l_n)$ le vecteur des réalisations des v.a. $L_1, \ldots, L_n$ et soit $\bar{\mathbf{l}}$ la moyenne empirique définie par : $\bar{\mathbf{l}} = \frac{1}{n}\sum_{i=1}^n l_i$.
- On peut alors "pivoter" g et résoudre en μ comme suit :

$$\begin{aligned}
&P(u(\tfrac{\alpha}{2}) \leq g(l_1, \ldots, l_n; \mu) \leq u(1 - \tfrac{\alpha}{2})) \\
&= P(u(\tfrac{\alpha}{2}) \leq \tfrac{\bar{\mathbf{l}} - \mu}{1/\sqrt{n}} \leq u(1 - \tfrac{\alpha}{2})) \\
&= P(\tfrac{u(\frac{\alpha}{2})}{\sqrt{n}} \leq \bar{\mathbf{l}} - \mu \leq \tfrac{u(1-\frac{\alpha}{2})}{\sqrt{n}}) \\
&= P(\tfrac{u(\frac{\alpha}{2})}{\sqrt{n}} - \bar{\mathbf{l}} \leq -\mu \leq \tfrac{u(1-\frac{\alpha}{2})}{\sqrt{n}} - \bar{\mathbf{l}}) \\
&= P(\bar{\mathbf{l}} - \tfrac{u(\frac{\alpha}{2})}{\sqrt{n}} \geq \mu \geq \bar{\mathbf{l}} - \tfrac{u(1-\frac{\alpha}{2})}{\sqrt{n}}) \\
&= P(\bar{\mathbf{l}} - \tfrac{u(1-\frac{\alpha}{2})}{\sqrt{n}} \leq \mu \leq \bar{\mathbf{l}} + \tfrac{u(1-\frac{\alpha}{2})}{\sqrt{n}}) \text{ car } u(\tfrac{\alpha}{2}) = -u(1 - \tfrac{\alpha}{2}).
\end{aligned}$$

- On a donc :

$$\mathbb{IC}_{1-\alpha}(\mu) = [\bar{\mathbf{l}} \mp \frac{u(1 - \frac{\alpha}{2})}{\sqrt{n}}]$$

Intervalles de confiance des estimateurs MCO

- Rappelons que sous $\mathcal{H}_{\mathbf{X}}$ et $\mathcal{H}_\epsilon$ et en supposant σ^2 inconnue, on a :

$$\forall j = 1, \ldots, p : T_j = \frac{a_j^* - a_j}{\hat{\sigma}^* \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} \sim \mathcal{T}_{n-(p+1)}$$

- Nous pouvons utiliser T_j telle une fonction pivot afin de déterminer un IC pour a_j.
- Pour tout coefficient $a_j, j = 1, \ldots, p$, nous avons alors l'IC bilatéral de niveau $1 - \alpha$ suivant :

$$\mathbb{IC}_{1-\alpha}(a_j) = \left[\hat{a}_j^* \mp t_{n-(p+1)}(1 - \alpha/2)\hat{\sigma}^* \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}\right]$$

où $t_m(\alpha)$ est le quantile de la loi de Student à m d.d.l. défini par $P(\mathcal{T}_m \leq t_m(\alpha)) = \alpha$.

Intervalles de confiance des estimateurs MCO (suite)

- Pour la variance résiduelle, rappelons que sous $\mathcal{H}_{\mathbf{X}}$ et $\mathcal{H}_\epsilon$, on a :

$$(n - (p+1))\frac{\sigma^{*2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

- De la même façon que précédemment, nous pouvons utiliser $(n - (p+1))\frac{\sigma^{*2}}{\sigma^2}$ comme fonction pivot pour trouver un IC de σ^2.
- Nous avons alors l'IC bilatéral de niveau $1 - \alpha$ suivant :

$$\mathbb{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n - (p+1))\hat{\sigma}^{*2}}{q_{n-(p+1)}(1 - \frac{\alpha}{2})}, \frac{(n - (p+1))\hat{\sigma}^{*2}}{q_{n-(p+1)}(\frac{\alpha}{2})}\right]$$

où $q_m(\alpha)$ est le quantile de la loi du Chi2 à m d.d.l. défini par $P(\chi^2_m \leq q_m(\alpha)) = \alpha$.

Prévisions et IC des prévisions

- Soit $\mathbf{x} = (1, x_1, \ldots, x_p)$ une observation quelconque de $\mathbb{R}^{p+1}$ (un individu non observé). On note $Y = \mathbf{x}^\top \mathbf{a} + \epsilon$, la variable cible associée.
- Y^* la v.a. de la prédiction obtenue par l'estimateur $\mathbf{a}^*$ vaut :

$$Y^* = \mathbf{x}^\top \mathbf{a}^* = \sum_{j=0}^{p} x_j a_j^*$$

- Avec $\mathcal{H}_\epsilon$ et les lois sur les estimateurs (supposant σ^2 connue), on a :

$$\begin{aligned}
\mathrm{E}_{Y|X}(Y - Y^*|\mathbf{X}) &= 0 \\
\mathrm{V}_{Y|X}(Y - Y^*|\mathbf{X}) &= \mathrm{V}_{Y|X}(\mathbf{x}^\top \mathbf{a} + \epsilon - \mathbf{x}^\top \mathbf{a}^*|\mathbf{X}) \\
&= \mathrm{V}_{Y|X}(\mathbf{x}^\top (\mathbf{a} - \mathbf{a}^*) + \epsilon|\mathbf{X}) \\
&= \mathrm{V}_{Y|X}(\mathbf{x}^\top (\mathbf{a} - \mathbf{a}^*)|\mathbf{X}) + \mathrm{V}(\epsilon) \text{ (car } \mathbf{a}^* \perp \epsilon) \\
&= \mathbf{x}^\top \mathrm{V}_{Y|X}(\mathbf{a}^* - \mathbf{a}|\mathbf{X})\mathbf{x} + \sigma^2 \text{ (car } \mathrm{V}(-Y) = \mathrm{V}(Y)) \\
&= \sigma^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x} + \sigma^2
\end{aligned}$$

Prévisions et IC des prévisions (suite)

- De ce qui précède nous en déduisons :

$$\frac{Y - Y^*}{\sqrt{\sigma^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x} + \sigma^2}} = \frac{Y - Y^*}{\sigma\sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x} + 1}} \sim \mathcal{N}(0, 1)$$

- **Mais**, σ^2 étant inconnue, il nous faut le remplacer par un estimateur σ^{*2} et on sait que $(n - (p+1))\frac{\sigma^{*2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
- Notons :
  - $U = \frac{Y - Y^*}{\sigma\sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}+1}} \sim \mathcal{N}(0, 1)$.
  - $V = (n - (p+1))\frac{\sigma^{*2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$
- Nous avons alors :

$$\frac{U}{\sqrt{V/(n - (p+1))}} = \frac{Y - Y^*}{\sigma^*\sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x} + 1}} \sim \mathcal{T}_{n-(p+1)}$$

- Nous avons alors pour Y l'IC bilatéral de niveau $1 - \alpha$ suivant :

$$\mathbb{IC}_{1-\alpha}(Y) = \left[\mathbf{x}^\top \hat{\mathbf{a}}^* \mp t_{n-(p+1)}(1 - \alpha/2)\hat{\sigma}^* \sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x} + 1}\right]$$

Tests d'hypothèses, exemple introductif

- Reprenons la régression linéaire multiple de l'exemple introductif.
- On obtient les coefficients estimés suivants :

```
> coefficients(lm(X$maxO3~X$T12+X$Vx+X$Ne12))
(Intercept)       X$T12         X$Vx       X$Ne12
 84.5473326   1.3150459   0.4864456  -4.8933729
```

- Est-ce que la variable à expliquer O_3 est influencée par la variable explicative VX ?
- Est-ce que la variable à expliquer O_3 est influencée par VX ou T_{12} ?

Tests d'hypothèses, exemple introductif (suite)

- Pour l'exemple précédent notre modèle est formellement défini par :

$$\underbrace{O_3}_{Y} = a_0 + a_1 \underbrace{T_{12}}_{X_1} + a_2 \underbrace{VX}_{X_2} + a_3 \underbrace{NE_{12}}_{X_3} + \epsilon$$

- Décider si VX influence O_3 revient à tester les hypothèses :

$$\mathcal{H}_0 : a_2 = 0 \text{ contre } \mathcal{H}_1 : a_2 \neq 0$$

- Décider si VX ou T_{12} influencent O_3 revient à tester :

$$\mathcal{H}_0 : a_1 = a_2 = 0 \text{ contre } \mathcal{H}_1 : a_1 \neq 0 \vee a_2 \neq 0$$

Rappels sur les tests d'hypothèses

- Un **test** est un procédé permettant de décider entre deux hypothèses à partir de la réalisation d'un échantillon aléatoire.
- Les hypothèses sont notées $\mathcal{H}_0$ et $\mathcal{H}_1$. Une et une seule de ces deux alternatives est vraie. $\mathcal{H}_0$ joue en général un rôle prédominant par rapport à $\mathcal{H}_1$.
- A l'issue d'une décision, il existe quatres situations possibles :

	$\mathcal{H}_0$ vraie	$\mathcal{H}_1$ vraie
$\mathcal{H}_0$ décidée	$1-\alpha$	β
$\mathcal{H}_1$ décidée	α	$1-\beta$

- On définit les concepts suivants :
  - L'**erreur de première espèce** associée au risque α : c'est rejeté $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie. α est alors la probabilité de rejeter à tort $\mathcal{H}_0$.
  - L'**erreur de deuxième espèce** associée au risque β : c'est accepté $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie. β est alors la probabilité d'accepter à tort $\mathcal{H}_0$.
  - La **puissance du test** est la valeur $1-\beta$ et est la probabilité de rejeter à raison $\mathcal{H}_0$.

Rappels sur les tests d'hypothèses (suite)

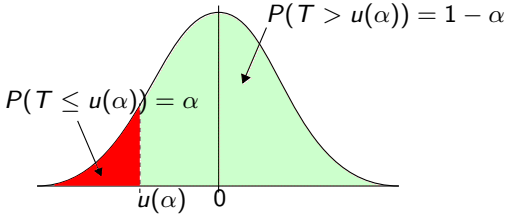
- La situation idéale serait que α et β soit nulle.
- On raisonnera le plus souvent avec l'erreur de première espèce.
- α est limitée à un niveau dit **niveau de significativité**. On prendra en général $\alpha = 10\%$ ou 5% ou 1%. On veut donc prendre peu de risque de rejeter à tort $\mathcal{H}_0$.
- Il existe deux types de tests :
  - Le **test bilatéral** : c'est quand on cherche une différence entre deux paramètres ou entre un paramètre et une valeur donnée sans se préoccuper du signe ou du sens de la différence. Exemple : $\mathcal{H}_0 : \theta = 0$ contre $\mathcal{H}_1 : \theta \neq 0$.
  - Le **test unilatéral** : c'est quand on cherche à savoir si un paramètre est supérieur ou inférieur à un autre ou à une valeur donnée. Exemple : $\mathcal{H}_0 : \theta > 0$ contre $\mathcal{H}_1 : \theta \leq 0$.

Rappels sur les tests d'hypothèses (suite)

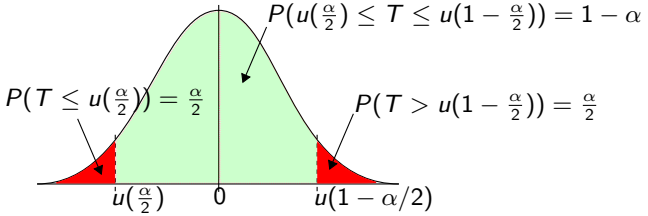
- Le risque de première espèce, α, étant fixé, il faut choisir une **variable de décision** appelée également **statistique de test** sur laquelle nous allons nous reposer pour choisir entre $\mathcal{H}_0$ et $\mathcal{H}_1$.
- Cette statistique est définie de manière à apporter une information sur le problème posé. Sa loi doit être connue dans le cadre d'au moins une des deux hypothèses. Le plus souvent, nous chercherons la **loi de la statistique de test en supposant que $\mathcal{H}_0$ est vraie**.
- La **région critique** appelée également **zone de rejet** et que l'on notera par $\mathbb{R}$, est l'ensemble des valeurs de la statistique de test qui conduit à rejeter $\mathcal{H}_0$ au profit de $\mathcal{H}_1$.

Rappels sur les tests d'hypothèses (suite)

- Exemple : Sous $\mathcal{H}_0$, la statistique de test T suit une $\mathcal{N}(0,1)$.
- Cas d'un test unilatéral pour une borne inférieure.



- Cas d'un test bilatéral symétrique.



Rappels sur les tests d'hypothèses (suite)

- La démarche d'un test est typiquement la suivante :
- Choix de $\mathcal{H}_0$ et de $\mathcal{H}_1$.
- Détermination de la statistique de test.
- Allure de la région de rejet :
  - Si c'est un test bilatéral : la zone de rejet de l'hypothèse principale se fait de part et d'autre de la distribution de probabilité de la statistique.
  - Si c'est un test unilatéral : la zone de rejet de l'hypothèse principale est située d'un seul côté de la distribution de probabilité de la statistique.
- Calcul de la région critique en fonction de α.
- Calcul de la valeur de la statistique de test à partir des observations de l'échantillon.
- Conclusion du test (est-ce qu'on est dans la région de rejet ?).

Test sur un coefficient a_j

- Etant donné un réel a ($a = 0$ en général), on veut tester :

$$\mathcal{H}_0 : a_j^* = a \text{ contre } \mathcal{H}_1 : a_j^* \neq a$$

- Sous $\mathcal{H}_0$, nous avons la propriété suivante :

$$T_j = \frac{a_j^* - a}{\hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim \mathcal{T}_{n-(p+1)} \text{ loi de Student à } n-(p+1) \text{ d.d.l.}$$

- La région de rejet de $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$ au niveau α est :

$$\mathbb{R}_\alpha = \{|\hat{T}_j| \geq t_{n-(p+1)}(1-\frac{\alpha}{2})\}$$

où $t_m(\alpha)$ est le quantile de la loi de Student à m d.d.l. défini par $P(\mathcal{T}_m \leq t_m(\alpha)) = \alpha$.

Test sur une combinaison linéaire de coefficients

- Etant donné un vecteur $\mathbf{b}$ de taille $(p \times 1)$ et un réel a, on veut tester :

$$\mathcal{H}_0 : \mathbf{b}^\top \mathbf{a}^* = a \text{ contre } \mathcal{H}_1 : \mathbf{b}^\top \mathbf{a}^* \neq a$$

- Sous $\mathcal{H}_0$, nous avons la propriété suivante :

$$T = \frac{\mathbf{b}^\top \mathbf{a}^* - a}{\hat{\sigma}\sqrt{\mathbf{b}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{b}}} \sim \mathcal{T}_{n-(p+1)}$$

- Comme précédemment, la région de rejet de $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$ au niveau α est :

$$\mathbb{R}_\alpha = \{|\hat{T}| \geq t_{n-(p+1)}(1 - \frac{\alpha}{2})\}$$

Exemple (code R)

```
> summary(lm(X$maxO3~X$T12+X$Vx+X$Ne12))

Call:
lm(formula = X$maxO3 ~ X$T12 + X$Vx + X$Ne12)

Residuals:
     Min       1Q   Median       3Q      Max
-29.0461  -8.4824   0.7861   7.7024  28.2916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.5473    13.6067   6.214 1.38e-07 ***
X$T12         1.3150     0.4974   2.644  0.01117 *
X$Vx          0.4864     0.1675   2.903  0.00565 **
X$Ne12       -4.8934     1.0270  -4.765 1.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

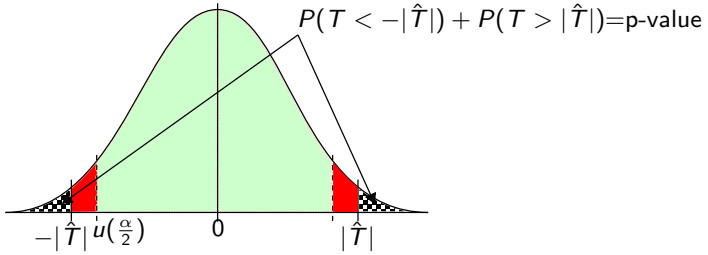

Exemple (code R) et interprétations

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.5473    13.6067   6.214 1.38e-07 ***
X$T12         1.3150     0.4974   2.644  0.01117 *
X$Vx          0.4864     0.1675   2.903  0.00565 **
X$Ne12       -4.8934     1.0270  -4.765 1.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- L'avant dernière colonne est connue sous le vocable de "**p-value**" qui est la probabilité d'observer une valeur de la statistique de test plus grande que la valeur estimée.
- Si la p-value est inférieure au risque de 1ère espéce α alors on rejette $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$.
- Graphiquement dans la dernière colonne cela correspond au signe comportant des $*$ (dans l'exemple on rejette tout le temps $\mathcal{H}_0$).

Rappels sur les tests d'hypothèses (suite)

- Représentation graphique de la p-value.
- Exemple : Sous $\mathcal{H}_0$, la statistique de test T suit une $\mathcal{N}(0, 1)$.
- Cas d'un test bilatéral symmétrique (ici on rejette $\mathcal{H}_0$).



- La p-value est la probabilité représentée en damier. Si elle est inférieure à α alors on rejette $\mathcal{H}_0$.

Test sur la variance résiduelle σ^2

- Etant donné une valeur σ_0, on veut tester :

$$\mathcal{H}_0 : \sigma = \sigma_0 \text{ contre } \mathcal{H}_1 : \sigma \neq \sigma_0$$

- Sous $\mathcal{H}_0$, nous avons la propriété suivante :

$$S^2 = (n-(p+1))\frac{\hat{\sigma}^2}{\sigma_0} \sim \chi^2_{n-(p+1)}$$

- La région de rejet de $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$ au niveau α est :

$$\mathbb{R}_\alpha = \{\hat{S}^2 \leq q_{n-(p+1)}(\frac{\alpha}{2})\} \cup \{S^2 \geq q_{n-(p+1)}(1-\frac{\alpha}{2})\}$$

où $q_m(\alpha)$ est le quantile de la loi du Chi2 à m d.d.l. défini par $P(\chi^2_m \leq q_m(\alpha)) = \alpha$.

Tests entre modèles emboîtés

- On veut tester la nullité d'un sous-ensemble de $q \leq p+1$ coefficients de $\mathbf{a}$. Cela revient à enlever q variables exogènes du modèle.
- Notons $\mathbb{A}_0 \subset \mathbb{A}$ le sous-ensemble des $p_0 = p+1-q$ variables exogènes restantes.
- On compare en fait deux modèles :

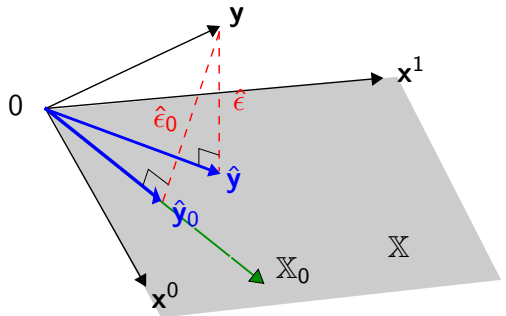
$$\begin{aligned} \mathbb{A}_0 &: Y|X \sim \mathcal{N}(\mathbf{X}_0^\top \mathbf{a}_0^*, \sigma^2) \\ \mathbb{A} &: Y|X \sim \mathcal{N}(\mathbf{X}^\top \mathbf{a}^*, \sigma^2) \end{aligned}$$

où $\mathbf{X}_0^\top$ est la matrice des données dont les colonnes sont les variables de $\mathbb{A}_0$ et $\mathbf{a}_0^*$ est le vecteur des estimateurs des MCO de taille $(p_0 \times 1)$.

- Cela revient à tester les hypothèses :

$$\mathcal{H}_0 : \forall j \in \mathbb{A} \setminus \mathbb{A}_0 : a_j^* = 0 \text{ contre } \mathcal{H}_1 : \exists j \in \mathbb{A} \setminus \mathbb{A}_0 : a_j^* \neq 0$$

Tests entre modèles emboîtés, interprétation géométrique



- On calcule $\hat{\mathbf{y}}_0$ et $\hat{\mathbf{y}}$ qui sont les projections orhtogonales sur les sev $\mathbb{X}_0$ et $\mathbb{X}$ engendrés respectivement par les colonnes de $\mathbf{X}_0$ et $\mathbf{X}$.
- Si $\hat{\mathbf{y}}_0$ et $\hat{\mathbf{y}}$ sont proches, on décide de ne pas rejeter $\mathcal{H}_0$ et donc de conserver le modèle le plus simple (principe de parcimonie).
- La statistique employée ici est donc fonction de la distance euclidienne entre les vecteurs aléatoires Y_0^* et Y^*.

Tests entre modèles emboîtés (suite)

- La statistique de test utilisée est :

$$F = \frac{\frac{1}{p+1-p_0}\|Y^* - Y_0^*\|^2}{\hat{\sigma}^{*2}}$$

où Y^* et Y_0^* sont les vecteurs aléatoires des prédictions par MCO basées sur $\mathbb{A}$ et $\mathbb{A}_0$ respectivement et où σ^{*2} est l'estimateur des MCO de la variance résiduelle de $\mathbb{A}$ (on divise par $\hat{\sigma}^{*2}$ pour supprimer les effets d'échelles).

- Notons :

$$\begin{aligned} scr(f) &= \sum_{i=1}^n (Y_i - Y_i^*)^2 = \|Y - Y^*\|^2 \\ scr(f_0) &= \sum_{i=1}^n (Y_i - Y_{0,i}^*)^2 = \|Y - Y_0^*\|^2 \end{aligned}$$

- On montre que :

$$F = \frac{(scr(f_0) - scr(f))/(p+1-p_0)}{scr(f)/(n-(p+1))}$$

Tests entre modèles emboîtés (suite)

- Nous avons le résultat suivant :

$$F = \frac{(scr(f_0) - scr(f))/(p+1-p_0)}{scr(f)/(n-(p+1))} \sim \mathcal{F}_{p+1-p_0,n-(p+1)}$$

où $\mathcal{F}_{m,s}$ est la loi de Fisher à m et s d.d.l.

- La région de rejet de $\mathcal{H}_0$ est la suivante :

$$\mathbb{R}_\alpha = \{\hat{F} \geq f_{p+1-p_0,n-(p+1)}(1-\alpha)\}$$

où $f_{m,s}(\alpha)$ est le quantile de la loi de Fisher défini par $P(\mathcal{F}_{m,s} \leq f_{m,s}(\alpha)) = \alpha$.

Tests de Fisher global

- Cas particulier du test précédent. On teste la nullité de tous les coefficients des variables exogènes exceptée la constante a_0.

$$\mathcal{H}_0 : \forall j \in \{1,\ldots,p\} : a_j = 0 \text{ contre } \mathcal{H}_1 : \exists j \in \{1,\ldots,p\} \ : a_j \neq 0$$

- Ce test est appelé **test de Fisher global**.
- Sous $\mathcal{H}_0$, on a : $p_0 = 1$ (dimension de l'espace $\mathbb{X}_0$), $a_0^* = \overline{Y}$ et $Y_0^* = \overline{Y}\mathbf{1}_n$ où $\mathbf{1}_n$ est le vecteur de taille $(n \times 1)$ rempli de 1.
- La statistique de test est alors :

$$F = \frac{(scr(f_0) - scr(f))/p}{scr(f)/(n-(p+1))} \sim \mathcal{F}_{p,n-(p+1)}$$

- La région de rejet de $\mathcal{H}_0$ est alors :

$$\mathbb{R}_\alpha = \{\hat{F} \geq f_{p,n-(p+1)}(1-\alpha)\}$$

Décomposition de la variance

- On distingue plusieurs types de variance pour l'analyse d'un ML $Y = f(X) = X^\top \mathbf{a} + \epsilon$.
- Soit Y_i^* la v.a. de la prédiction de X_i obtenue par l'estimateur des MCO. On définit les trois types de variances suivantes.
- **Variance résiduelle** $(scr(f)/n)$: $\frac{1}{n}\sum_{i=1}^n (Y_i - Y_i^*)^2$.
- **Variance expliquée** $(sce(f)/n)$: $\frac{1}{n}\sum_{i=1}^n (Y_i^* - \overline{Y})^2$.
- **Variance totale** (sct/n) : $\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$.

Variance	Somme des carrés	d.d.l.
Résiduelle	$scr(f) = \|Y - Y^*\|^2$	$n-(p+1)$
Expliquée	$sce(f) = \|Y^* - \overline{Y}\mathbf{1}_n\|^2$	p
Totale	$sct = \|Y - \overline{Y}\mathbf{1}_n\|^2$	$n-1$

- **Décomposition de la variance**. On montre que :

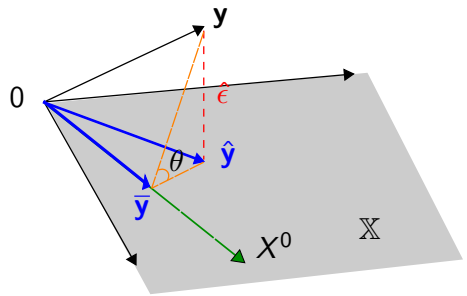
$$sct = sce(f) + scr(f)$$

Tests de Fisher global et le R^2

- Un autre outil pour évaluer un ML est le **coefficient de corrélation (ou de détermination) multiple** dénoté R^2, défini par :

$$R^2 = \frac{sce(f)}{sct}$$

- $R^2 = (\cos\theta)^2$ où θ est l'angle formé par les vecteurs Y et Y^* recentrés par rapport au vecteur moyen $\overline{Y}$.



- $0 \leq R^2 \leq 1$
- Plus R^2 est proche de 1, meilleur est l'ajustement.

Tests de Fisher global et le R^2

- On montre que le test de Fisher global est lié au R^2 puisque :

$$F = \frac{n-(p+1)}{p}\frac{R^2}{1-R^2}$$

- Le R^2 ne tient pas compte du nombre de variables explicatives.
- Afin de pallier à ce défaut pour comparer deux (ou plusieurs) modèles $\mathbb{A}_0$ et $\mathbb{A}$ n'ayant pas le même nombre de variables on utilise le coefficient de corrélation (ou de détermination) **ajusté** défini par :

$$R_a^2 = 1 - \frac{scr(f)/(n-(p+1))}{sct/(n-1)}$$

Exemple (code R)

```
> summary(lm(X$maxO3~X$T12+X$Vx+X$Ne12))

Call:
lm(formula = X$maxO3 ~ X$T12 + X$Vx + X$Ne12)

Residuals:
     Min       1Q   Median       3Q      Max
-29.0461  -8.4824   0.7861   7.7024  28.2916

...

Residual standard error: 13.91 on 46 degrees of freedom
Multiple R-squared: 0.6819,Adjusted R-squared: 0.6611
F-statistic: 32.87 on 3 and 46 DF,  p-value: 1.663e-11
```

- Dans les dernières lignes nous avons le résultat du test de Fisher global et le R_a^2.

Rappel du Sommaire

Introduction

- En régression linéaire, les grandes étapes de la démarche sont :
  - Formulation du problème $Y = X^\top \mathbf{a} + \epsilon$.
  - Estimation du modèle par MV et inférence dans le modèle gaussien.
  - **Validation des hypothèses** sous-jacentes au modèle gaussien.
- L'étape de validation des hypothèses est importante car si celles-ci ne sont pas vérifiées, alors les tests et IC ne sont plus valides.
- Rappelons les hypothèses du modèle linéaire gaussien :
  - Linéarité du modèle :
    - $\forall i = 1, \ldots, n : Y_i = X_i^\top \mathbf{a} + \epsilon_i$.
  - $\mathcal{H}_{\mathbf{X}}$:
    - La matrice de données $\mathbf{X}$ est de plein rang, $rg(\mathbf{X}) = p$.
  - $\mathcal{H}_\epsilon$:
    - Moyenne nulle, $\forall i : \mathrm{E}(\epsilon_i) = 0$.
    - Homoscédasticité, $\forall i : \mathrm{V}(\epsilon_i) = \sigma^2$.
    - Non corrélation, $\forall i \neq j : \mathrm{C}(\epsilon_i, \epsilon_j) = 0$.
    - Normalité, $\forall i : \epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Analyse des résidus

- Nous cherchons ici à **valider l'hypothèse de linéarité et** $\mathcal{H}_\epsilon$ (nous aborderons plus loin le cas de $\mathcal{H}_{\mathbf{X}}$).
- La validation de l'hypothèse de linéarité et de $\mathcal{H}_\epsilon$ repose sur l'**analyse des résidus empiriques**.
- Rappelons les quantités en jeu ici :
  - Résidus théoriques, $\forall i : \epsilon_i$.
  - Résidus empiriques (estimateurs des ϵ_i), $\forall i : \epsilon_i^* = Y_i - Y_i^*$
- Nous avons par ailleurs les résultats suivants par construction :
  - Moyenne nulle des résidus empirique : $\mathrm{E}(\epsilon^*) = \mathbf{0}$.
  - Variance des résidus empirique : $\mathrm{V}(\epsilon^*) = \sigma^2(\mathbf{I}_n - \mathbf{P}_{\mathbb{X}})$ (cf slide 93).
- On va utiliser des méthodes graphiques pour analyser les hypothèses.

Analyse des résidus / Données aberrantes

- Ici on cherche à **valider l'hypothèse de linéarité**.
- Dans un premier temps, des graphiques bidimensionnels entre $\mathbf{y}$ et chaque $\mathbf{x}^j$ permettent d'apprécier la plus ou moins grande dépendance linéaire entre la variable à expliquer et chaque variable explicative. On pourra d'ailleurs procéder à des **transformations des variables** (logarithme, puissance, ...) pour favoriser la dépendance linéaire.
- Dans un secont temps, l'**analyse des erreurs** comises permettent d'apprécier le plus ou moins bon ajustement des données par un modèle linéaire. S'il y a beaucoup d'erreurs importantes, l'hypothèse de linéarité doit être remise en cause.
- La détection d'**observations aberrantes** s'effectue également dans le cadre de cette analyse.

Analyse des résidus / Données aberrantes (suite)

- On analyse les **résidus studentisés** par **validation croisée** définis par, $\forall i = 1, \ldots, n$:

$$t_i^s = \frac{y_i - \mathbf{x}_i^\top \mathbf{a}_{(i)}^*}{\sigma_{(i)}^* \sqrt{1 - [\mathbf{P}_{\mathbb{X}}]_{ii}}}$$

où $\mathbf{a}_{(i)}^*$ et $\sigma_{(i)}^*$ sont les estimateurs des MCO **sans tenir compte de l'observation** $\mathbf{x}_i$.

- Sous $\mathcal{H}_\epsilon$, on montre que $\forall i = 1, \ldots, n$:

$$t_i^s \sim \mathcal{T}_{(n-1)-(p+1)}$$

- Ce résultat théorique nous permet d'avoir une méthode pratique pour décider si des observations sont aberrantes.

Analyse des résidus / Données aberrantes (suite)

- En effet, les t_i^s doivent être dans une bande de frontière ∓ 2. Cette valeur est proche du quantile de la loi de student de niveau 0.975.
- Les t_i^s qui sortent de cette bande correspondent aux observations dont les erreurs sont importantes.
- **Attention !** En théorie, 2.5% des points peuvent sortir de la bande. Donc, uniquement ceux qui sortent très largement peuvent être considérées comme **données aberrantes**.
- On dira que $(\mathbf{x}_i, y_i)$ est une **donnée aberrante si son résidu studentisé** $\hat{t}_i^s$ **est très élevé** :

$$|\hat{t}_i^s| > t_{(n-1)-(p+1)}(1 - \frac{\alpha}{2}) \quad \text{où } \alpha < 2.5\%.$$

- Les données aberrantes doivent être analysées pour comprendre d'où vient l'aberration (erreurs de mesures, données rares, ...). A l'issue de cette analyse, il peut être décidé d'enlever la ou les données aberrantes car celles-ci influent sur l'estimation des paramètres.

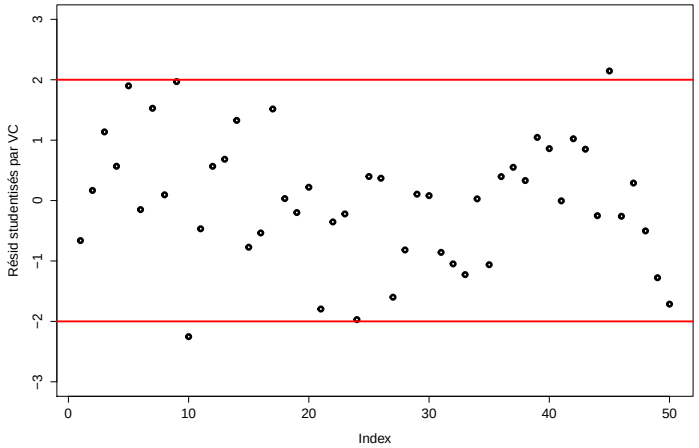
Analyse des résidus / Données aberrantes (code R)

```
> #Estimation du modèle
> reg_mult=lm(X$maxO3~X$T12+X$Vx+X$Ne12)
>
> #Calcul des résidus studentisés
> rstudent(reg_mult)
           1            2            3            4            5            6
-0.662059616  0.167746427  1.137022579  0.568170111  1.899178588 -0.149242455
           7            8            9           10           11           12
 1.527569304  0.095567916  1.969269151 -2.252622132 -0.467207922  0.567109301

...

>
> #Visualisation des résidus studentisés
> plot(rstudent(reg_mult),ylab="Résid studentisés par VC",lwd=4,ylim=c(-3,3))
> abline(h=c(-2,2),col="red",lwd=3)
```


Analyse des résidus / Données aberrantes (code R)



Analyse des résidus / Homoscédasticité

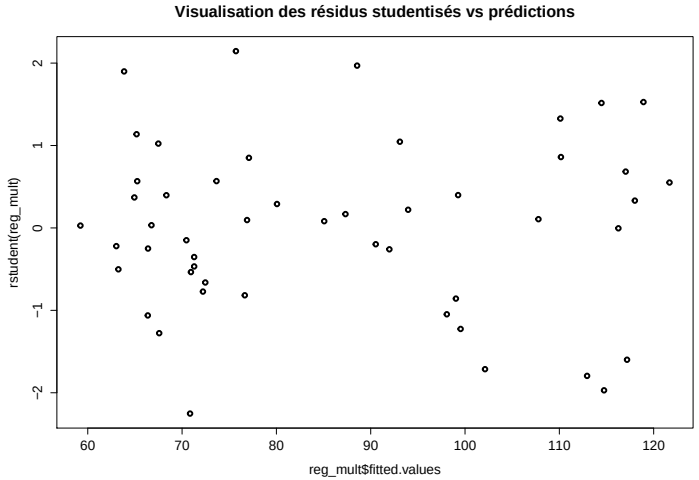
- Ici on cherche à **vérifier l'homogénéité de la variance résiduelle** :

$$\forall i = 1, \ldots, n : \mathrm{V}(\epsilon_i^*) = \sigma^{*2}$$

- Une méthode graphique consiste à **tracer le nuage de points** $\{(\hat{y}_i, \hat{t}_i^s)\}_{i=1}^n$.
- En fonction de la nature des données, on pourra prendre en abscisse d'autres variables à la place de $\hat{y}_i$ (comme le temps si les données sont temporelles).
- Si une structure apparaît (tendance, cône, vague, ...), l'hypothèse d'homoscédasticité risque de ne pas être vérifiée.
- Au-delà de l'approche graphique, des tests statistiques existent, pour plus de détails cf [Guyon (2011)].

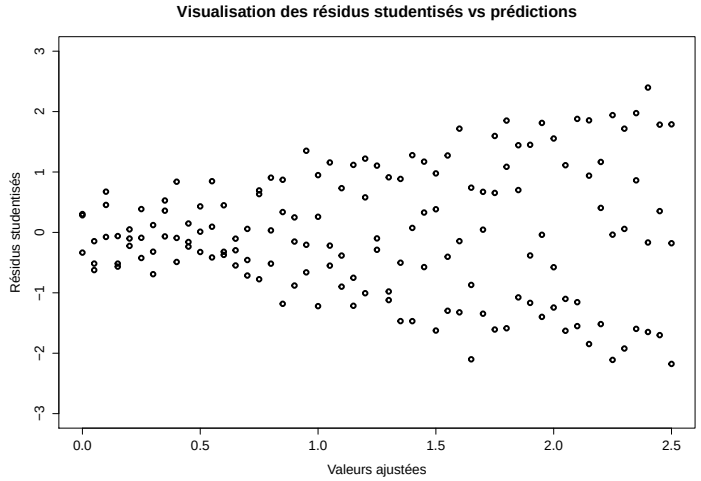
Analyse des résidus / Homoscédasticité (code R)

```
> #Visualisation des residus studentisés contre les prévisions
> plot(reg_mult$fitted.values,rstudent(reg_mult),main="Visualisation des résidu
```



Analyse des résidus / Homoscédasticité (code R)

- Pour des données indexés par le temps ou séquentielles, on peut avoir un cas d'hétéroscédasticité que l'on visualise par un "cône".



Analyse des résidus / Corrélation entre résidus

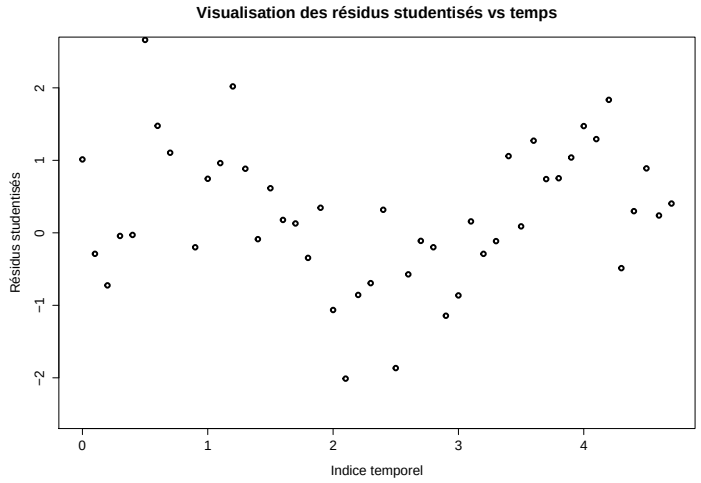
- Ici on cherche à **vérifier si les résidus sont corrélés ou non** :

$$\forall i, j = 1, \ldots, n : \mathrm{C}(\epsilon_i^*, \epsilon_j^*) = 0$$

- Ces corrélations peuvent être dues à plusieurs facteurs :
  - **Mauvaise modélisation** : supposons que X^k est une variable permettant d'expliquer Y mais que celle-ci n'est pas prise en considération dans le modèle. Dans ce cas, l'absence de X^k ne permet pas une bonne modélisation de Y et ce manque est répercuté au niveau des résidus. Pour observer graphiquement cette corrélation, on pourra tracer le nuage de points $\{(x_{ik}, \hat{t}_i^s)\}_{i=1}^n$ et détecter une structure
  - **Structuration temporelle** : si les données sont temporelles ou séquentielles (ie $\forall i$, l'observation de $\mathbf{x}_{i-1}$ précède celle de $\mathbf{x}_i$), il peut avoir une autocorrélation des résidus (ie $\epsilon_i = f(\epsilon_{i-1}) + \eta$). Si on suspecte un modèle autorégressif d'ordre 1 (AR) on utilise le test de Durbin-Watson où $\mathcal{H}_0$: indépendance entre ϵ_{i-1} et ϵ_i contre $\mathcal{H}_1 : \epsilon_i = \rho\epsilon_{i-1} + \eta_i$ où les η_i sont i.i.d. selon $\mathcal{N}(0, 1)$ et $\rho \neq 0$. On pourra observer le nuage de points $\{(i, \hat{t}_i^s)\}_{i=1}^n$ pour suspecter une autocorrélation des résidus.

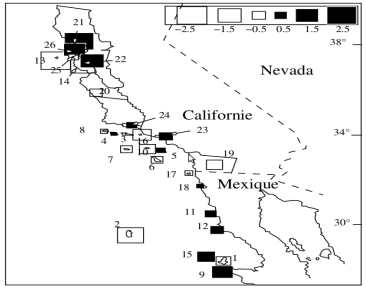
Analyse des résidus / Corrélation entre résidus (suite)

- Exemple d'autocorrélation des erreurs (structure de type "vague").



Analyse des résidus / Corrélation entre résidus (suite)

- Ces corrélations peuvent être dues à plusieurs facteurs (suite) :
  - **Structuration spatiale** : si les observations sont géo-référencées, il peut exister une structuration spatiale. Dans ce cas, l'observation des résidus studentisés (ou de leur valeur absolue) sur une carte pourra permettre de détecter une telle dépendance.
- Exemple tiré de [Cornillon et al (2011)] : Nombre de plantes endémiques en fonction de la surface de l'unité de mesure, l'altitude et la latitude. Résidus studentisés représentés sur une carte :



Analyse des résidus / Normalité

- Ici on cherche à **vérifier si les résidus suivent une loi normale ou pas**.
- On peut utiliser également une méthode graphique appelée **droite de Henry** ou graphique Quantile-Quantile ("**Q-Q plot**").
- Supposons que $\mathcal{H}_0$: $(Z_1, \ldots, Z_n)$ est un échantillon i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.
- Pour tester $\mathcal{H}_0$ on peut alors procéder de la façon suivante :
  - On réduit (ou standardise) l'échantillon : $\forall i = 1, \ldots, n : t_i = Z_i/\sigma$.
  - On classe les valeurs $(t_1, \ldots, t_n)$ dans l'ordre croissant. On a alors une permutation τ de l'ensemble $\{1, 2, \ldots, n\}$ telle que :

$$t_{\tau(1)} \leq t_{\tau(2)} \leq \ldots \leq t_{\tau(n)}$$

  - On a la propriété suivante : $\mathrm{E}(t_{\tau(i)}) \approx \phi^{-1}(\frac{2i-1}{2n})$ où ϕ est la fonction de répartition de $\mathcal{N}(0, 1)$.
  - On trace le nuage de points $\{(\phi^{-1}(\frac{2i-1}{2n}), t_{\tau(i)})\}_{i=1}^n$.
- Sous $\mathcal{H}_0$ les points $\{(\phi^{-1}(\frac{2i-1}{2n}), t_{\tau(i)})\}_{i=1}^n$ sont approximativement alignés selon la 1ère bissectrice.

Analyse des résidus / Normalité (suite)

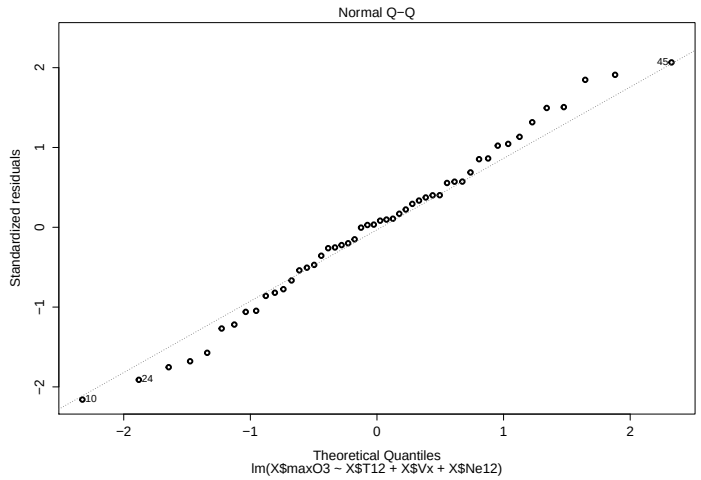
- On applique ce principe aux résidus $(\hat{\epsilon}_1^*, \hat{\epsilon}_2^*, \ldots, \hat{\epsilon}_n^*)$:
  - On réduit l'échantillon : $\forall i = 1, \ldots, n : \hat{t}_i = \hat{\epsilon}_i^*/\hat{\sigma}$.
  - On détermine la permutation τ telle que :

$$\hat{t}_{\tau(1)} \leq \hat{t}_{\tau(2)} \leq \ldots \leq \hat{t}_{\tau(n)}$$

  - On trace le nuage de points $\{(\phi^{-1}(\frac{2i-1}{2n}), \hat{t}_{\tau(i)})\}_{i=1}^n$.
- Plus le nuage de points est aligné selon la 1ère bissectrice plus on accepte l'hypothèse de normalité.

Analyse des résidus / Normalité (code R)

```
> #Visualisation du QQ plot
> plot(reg_mult,which=2,lwd=3)
```



Conclusions sur l'analyse des résidus

- L'analyse des résidus empiriques permet de valider les hypothèses du modèle linaire gaussien.
- Des méthodes graphiques simples peuvent être mises en oeuvre :
  - Le graphe des $\{(i, \hat{t}_i^s)\}$ permet de voir la qualité de l'ajustement et de détecter les points aberrants.
  - Le graphe des $\{(\hat{y}_i, \hat{t}_i^s)\}$ permet de voir si la variance résiduelle est constante ou non.
  - Pour des données séquentielles, le graphe des $\{(i, \hat{t}_i^s)\}$ permet de voir s'il existe des autocorrélations entre résidus.
  - Pour des données spatiales, la représentation des $\{\hat{t}_i^s\}$ sur une carte permet de voir s'il existe des corrélations dues au géo-référencement.
- Comme déjà mentionné, **des tests statistiques peuvent accompagner l'analyse graphique** avant de valider ou non les hypothèses.
- Pour aller plus loin dans la validation d'un modèle linéaire gaussien on peut également étudier la **robustesse des estimations** : que se passe t-il si on enlève une observation ? Les estimations changent-elles beaucoup ?

Rappel du Sommaire

Introduction

- Nous avons considéré jusqu'à présent que notre modèle de régression linéaire était valide, $\forall i$:

$$Y_i = X_i^\top \mathbf{a} + \epsilon_i$$

- Nous avons de plus fait parmi $\mathcal{H}_\epsilon$, l'hypothèse suivante pour le vecteur des résidus $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$:

$$\mathrm{V}(\epsilon) = \sigma^2 \mathbf{I}_n$$

- Il existe souvent des cas ou cette hypothèse n'est pas satisfaite. Deux situations peuvent se produire :
  - La **variance n'est pas constante** (**hétéroscédasticité**) :

$$\forall i \neq j : \sigma_i^2 \neq \sigma_j^2$$

  - Les **résidus sont corrélés entre eux** :

$$\forall i \neq j : \mathrm{C}(\epsilon_i, \epsilon_j) \neq 0$$

Cas général de la matrice de variance-covariance

- Afin de tenir compte de ces cas plus généraux, nous noterons la **matrice de variance-covariance** :

$$\mathbf{\Sigma}_\epsilon = \sigma^2 \mathbf{\Omega}$$

où $\mathbf{\Omega}$ est une matrice carrée d'ordre n.

- Nous faisons les hypothèses suivantes dénotées $\mathcal{H}_{\mathbf{\Omega}}$:
  - $\mathbf{\Omega}$ est symmétrique : $\mathbf{\Omega}^\top = \mathbf{\Omega}$
  - $\mathbf{\Omega}$ est définie positivite : $\forall \mathbf{x} \in \mathbb{R}^p$: $\mathbf{x}^\top \mathbf{\Omega} \mathbf{x} \geq 0$ et $\mathbf{x}^\top \mathbf{\Omega} \mathbf{x} = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ (les valeurs propres de $\mathbf{\Omega}$ sont toutes strictement positives).
  - $\mathbf{\Omega}$ est de plein rang : $rg(\mathbf{\Omega}) = n$ ($det(\mathbf{\Omega}) \neq 0$).
- Supposons pour l'instant que σ^2 **et** $\mathbf{\Omega}$ **sont connus**. Quel impact cela a t-il sur les estimateurs des MCO ou (MV) ?

Estimateurs des MCO

- A la place de $\mathcal{H}_\epsilon$ on suppose $\mathcal{H}_{\mathbf{\Omega}}$ avec σ^2 et $\mathbf{\Omega}$ donnés.
- L'estimateur des MCO reste bien sûr :

$$\mathbf{a}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

- Cet estimateur reste sans biais :

$$\mathrm{E}_Y(\mathbf{a}^* | \mathbf{X}) = \mathbf{a}$$

- Mais **sous** $\mathcal{H}_{\mathbf{\Omega}}$ **la variance diffère de précédemment** (cf slide 90) :

$$\begin{aligned} \mathrm{V}_Y(\mathbf{a}^* | \mathbf{X}) &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathrm{V}_Y(Y|\mathbf{X}) \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \\ &= \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{\Omega} \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \end{aligned}$$

- L'estimateur n'est plus celui de variance minimale parmi ceux qui sont sans biais (**le théorème de Gauss-Markov n'est plus vrai**).

Estimateurs des MCO (suite)

- Si on suppose cette fois-ci que σ^2 **n'est pas connu**, alors l'estimateur de la variance résiduelle dépend également de $\mathbf{\Omega}$ (cf slide 93) :

$$\begin{aligned}
\mathrm{E}(\|\epsilon\|^2) &= \mathrm{E}(\epsilon^\top \epsilon) \\
&= \mathrm{E}((\mathbf{P}_{\mathbb{X}^\perp}\epsilon)^\top \mathbf{P}_{\mathbb{X}^\perp}\epsilon) \\
&= \mathrm{E}(\epsilon^\top \mathbf{P}_{\mathbb{X}^\perp}\epsilon) \\
&= \mathrm{E}(\sum_{i,j=1}^{n} \epsilon_i \epsilon_j [\mathbf{P}_{\mathbb{X}^\perp}]_{ij}) \\
&= \sum_{i,j=1}^{n} \mathrm{E}(\epsilon_i \epsilon_j)[\mathbf{P}_{\mathbb{X}^\perp}]_{ij} \\
&= \sigma^2 \sum_{i,j=1}^{n} \mathbf{\Omega}_{ij}[\mathbf{P}_{\mathbb{X}^\perp}]_{ij}
\end{aligned}$$

- Ici également, on obtient un **estimateur biaisé de la variance résiduelle**

Limites des estimateurs des MCO

- Les **estimateurs des MCO sont donc "inadaptés"** (ie vis à vis du théorème de Gauss-Markov) dans le cas général (ie en faisant l'hypothèse $\mathcal{H}_{\mathbf{\Omega}}$).
- Nous allons donc proposer des estimateurs plus adéquats.
- Nous considérons dans un premier temps, le **cas de l'hétéroscédasticité** uniquement. Nous avons donc le modèle, $\forall i$:

$$Y_i = X_i^\top \mathbf{a} + \epsilon_i$$

avec les hypothèses sur le vecteur des résidus suivantes :
  - Moyenne nulle, $\forall i : \mathrm{E}(\epsilon_i) = 0$.
  - **Hétéroscédasticité**, $\forall i : \mathrm{V}(\epsilon_i) = \sigma^2\omega_i^2$.
  - Non corrélation, $\forall i \neq j : \mathrm{C}(\epsilon_i, \epsilon_j) = 0$.
- Cela revient à faire l'hypothèse que $\mathbf{\Omega}$ est diagonale :

$$\mathbf{\Omega} = \begin{pmatrix} \omega_1^2 & & \\ & \ddots & \\ & & \omega_n^2 \end{pmatrix}$$

Estimateurs des MCP (Moindres Carrés Pondérés)

- Sous l'hypothèse précédente, on voit que pour obtenir des estimateurs sans biais, **il faut se ramener à un cas homoscédastique** et donc homogénéiser la variance résiduelle issue de chaque observation.
- Pour cela, il suffit d'appliquer les **transformations** suivantes, $\forall i$:

$$\frac{y_i}{\omega_i} = a_0\frac{x_{i0}}{\omega_i} + a_1\frac{x_{i1}}{\omega_i} + \ldots + a_p\frac{x_{ip}}{\omega_i} + \frac{\epsilon_i}{\omega_i}$$

- Définissons alors les variables suivantes :

$$\forall i \ : \tilde{Y}_i = \frac{Y_i}{\omega_i} \quad ; \quad \forall i : \tilde{X}_i = \frac{X_i}{\omega_i} \quad \text{et} \quad \forall i : \tilde{\epsilon}_i = \frac{\epsilon_i}{\omega_i}$$

- Nous nous ramenons donc au modèle linéaire suivant, $\forall i$:

$$\tilde{Y}_i = \tilde{X}_i^\top \mathbf{a} + \tilde{\epsilon}_i$$

- Nous avons un **modèle transformé** qui **est homoscédastique !**
$\forall i \ : \mathrm{V}(\tilde{\epsilon}_i) = \mathrm{V}(\epsilon_i/\omega_i) = \mathrm{V}(\epsilon_i)/\omega_i^2 = \sigma^2\omega_i^2/\omega_i^2 = \sigma^2$

Estimateurs des MCP (suite)

- On peut utiliser les **MCO avec le modèle transformé** afin de retrouver toutes les bonnes propriétés. Pour ce faire, écrivons le modèle en termes matriciels.
- La matrice $\mathbf{\Omega}$ étant diagonale on peut déterminer facilement sa racine carré qui est telle que $\mathbf{\Omega}^{1/2}\mathbf{\Omega}^{1/2} = \mathbf{\Omega}$. En effet, on voit que $\mathbf{\Omega}^{1/2}$ est la matrice diagonale suivante :

$$\mathbf{\Omega}^{1/2} = \begin{pmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{pmatrix}$$

- Puis l'inverse $\mathbf{\Omega}^{-1/2}$ est telle que $\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{1/2} = \mathbf{I}_n$. On voit donc que :

$$\mathbf{\Omega}^{-1/2} = \begin{pmatrix} \frac{1}{\omega_1} & & \\ & \ddots & \\ & & \frac{1}{\omega_n} \end{pmatrix}$$

Estimateurs des MCP (suite)

- Les transformations précédentes appliquées aux observations, peuvent alors être définies matriciellement comme suit :

$$\tilde{\mathbf{y}} = \mathbf{\Omega}^{-1/2}\mathbf{y} \quad ; \quad \tilde{\mathbf{X}} = \mathbf{\Omega}^{-1/2}\mathbf{X}$$

- Nous pouvons appliquer au modèle transformé les MCO et on a donc :

$$\begin{aligned}
\hat{\mathbf{a}}_{mcp} &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}} \\
&= ((\mathbf{\Omega}^{-1/2}\mathbf{X})^\top\mathbf{\Omega}^{-1/2}\mathbf{X})^{-1}(\mathbf{\Omega}^{-1/2}\mathbf{X})^\top\mathbf{\Omega}^{-1/2}\mathbf{y} \\
&= (\mathbf{X}^\top\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{-1/2}\mathbf{y} \\
&= (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{y}
\end{aligned}$$

- Dans ce cas **le théorème de Gauss-Markov est à nouveau vérifié** : l'estimateur $\mathbf{a}_{mcp}$ dit des **moindres carrés pondérés** est parmi les estimateurs linéaires sans biais celui de variance minimale.

Estimateurs des MCG (Moindres Carrés Généralisés)

- Nous supposons le modèle suivant, $\forall i$:

$$Y_i = X_i^\top\mathbf{a} + \epsilon_i$$

avec les hypothèses suivantes sur le vecteur des résidus suivantes :
  - Moyenne nulle, $\forall i : \mathrm{E}(\epsilon_i) = 0$.
  - Matrice de variance-covariance hétéroscédastique et avec corrélations, $\mathrm{V}(\epsilon) = \sigma^2\mathbf{\Omega}$ avec $rg(\mathbf{\Omega}) = n$.
- Rappelons que $\mathbf{\Omega}$ est une matrice carrée d'ordre n et que $(\mathcal{H}_{\mathbf{\Omega}})$:
  - $\mathbf{\Omega}$ est symmétrique.
  - $\mathbf{\Omega}$ est définie positivite.
  - $\mathbf{\Omega}$ est de plein rang.
- Nous généralisons le cas précédent en adaptant une tranformation plus générale que celle associée aux MCP. En effet, contrairement aux MCP, on considère désormais le cas général où $\mathbf{\Omega}$ n'est pas diagonale.

Estimateurs des MCG (suite)

- $\mathbf{\Omega}$ étant réelle, symmétrique et définie positive, il existe une **décomposition spectrale** de la forme suivante :

$$\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}^{-1}$$

où $\mathbf{D}$ est la matrice diagonale formée des valeurs propres (strictement positives) et $\mathbf{\Lambda}$ de taille $(n \times n)$ est la matrice de passage formées des vecteurs propres normés (ie de norme unitaire).
- Par ailleurs, $\mathbf{\Omega}$ étant réelle symmétrique, nous avons la propriété que les vecteurs propres sont orthogonaux deux à deux. Ainsi :

$$\mathbf{\Lambda}^\top\mathbf{\Lambda} = \mathbf{I}_n$$

- Ceci implique que $\mathbf{\Lambda}$ est une **matrice orthogonale** :

$$\mathbf{\Lambda}^\top = \mathbf{\Lambda}^{-1}$$

- Posons $\mathbf{P} = \mathbf{\Lambda}\mathbf{D}^{1/2}$ (qui est inversible). Nous pouvons donc écrire :

$$\mathbf{\Omega} = \mathbf{P}\mathbf{P}^\top$$

Estimateurs des MCG (suite)

- Définissons alors les vecteurs suivants :

$$\tilde{Y} = \mathbf{P}^{-1}Y \quad ; \quad \forall i : \tilde{X}_i = \mathbf{P}^{-1}X_i \quad \text{et} \quad \tilde{\epsilon} = \mathbf{P}^{-1}\epsilon$$

- Nous posons alors, $\forall i$:

$$\tilde{Y}_i = \tilde{X}_i^\top\mathbf{a} + \tilde{\epsilon}_i$$

- Grâce à ces transformations, **on retrouve des hypothèses classiques** concernant les résidus $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)$ puisque nous avons :
  - Moyenne nulle, $\forall i : \mathrm{E}(\tilde{\epsilon}_i) = 0$.
  - Matrice de variance-covariance homoscédastique et sans corrélation !, $\mathrm{V}(\tilde{\epsilon}) = \sigma^2\mathbf{I}_n$.

Estimateurs des MCG (suite)

- Concernant les observations, nous avons donc les transformations suivantes :

$$\tilde{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{y} \quad ; \quad \tilde{\mathbf{X}} = \mathbf{P}^{-1}\mathbf{X}$$

- Nous pouvons donc utiliser les MCO avec le modèle transformé et retrouver toutes les bonnes propriétés des estimateurs. On définit alors **l'estimateur des moindres carrés généralisés (MCG)** que l'on nomme également **estimateur d'Aitken** :

$$\begin{aligned} \hat{\mathbf{a}}_{mcg} &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}} \\ &= (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y} \end{aligned}$$

Interprétation géométrique

- Dans $\mathbb{R}^n$ on utilise classiquement le produit scalaire canonique :

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top\mathbf{v}$$

- Mais il existe une infinité de produits scalaires dans un espace vectoriel (on parle également de métriques).
- La matrice $\mathbf{A}$ représente un produit scalaire si elle est symétrique et définie positive et on a :

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{A}} = \mathbf{u}^\top\mathbf{A}\mathbf{v}$$

- La matrice $\boldsymbol{\Omega}$ est symétrique et définie positive. On a la propriété que $\boldsymbol{\Omega}^{-1}$ est aussi une matrice symmétrique et définie positive.
- Donc $\boldsymbol{\Omega}^{-1}$ peut être interprétée telle un produit scalaire de $\mathbb{R}^n$:

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\boldsymbol{\Omega}^{-1}}$$

Interprétation géométrique (suite)

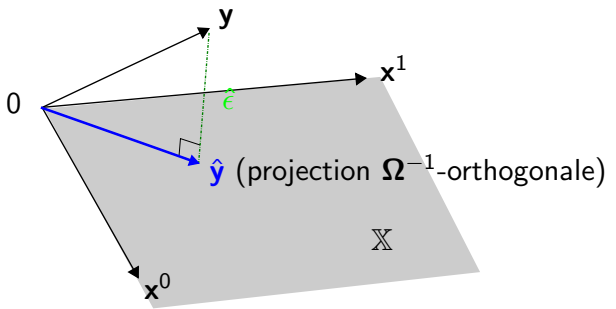
- L'estimateur des MCG peut être interprétée comme la solution du problème suivant :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{a}\|_{\boldsymbol{\Omega}^{-1}}$$

- Il est donné par la **projection dite $\boldsymbol{\Omega}^{-1}$-orthogonale sur le sous-espace de $\mathbb{R}^{p+1}$ engendré par les colonnes de $\mathbf{X}$**.
- Le vecteur des résidus ϵ est $\boldsymbol{\Omega}^{-1}$-orthogonal à tout vecteur appartenant à l'espace engendré par les colonnes de $\mathbf{X}$ et qui s'écrivent $\mathbf{v} = \mathbf{X}\mathbf{b}$ où $\mathbf{b} \in \mathbb{R}^{p+1}$:

$$\begin{aligned} \langle \mathbf{X}\mathbf{b}, \epsilon \rangle_{\boldsymbol{\Omega}^{-1}} = 0 &\Leftrightarrow \langle \mathbf{X}\mathbf{b}, Y - \mathbf{X}\mathbf{a}_{mcg} \rangle_{\boldsymbol{\Omega}^{-1}} = 0 \\ &\Leftrightarrow (\mathbf{X}\mathbf{b})^\top\boldsymbol{\Omega}^{-1}(Y - \mathbf{X}\mathbf{a}_{mcg}) = 0 \\ &\Leftrightarrow \mathbf{b}^\top\mathbf{X}^\top\boldsymbol{\Omega}^{-1}Y - \mathbf{b}^\top\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}\mathbf{a}_{mcg} = 0 \\ &\Leftrightarrow \mathbf{X}^\top\boldsymbol{\Omega}^{-1}Y = \mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}\mathbf{a}_{mcg} \\ &\Leftrightarrow (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}Y = \mathbf{a}_{mcg} \end{aligned}$$

Interprétations géométriques des MCG



- Les MCG consistent à projeter $\mathbf{y}$ $\boldsymbol{\Omega}^{-1}$-orthogonalement sur $\mathbb{X}$, le sous-espace vectoriel (sev) de $\mathbb{R}^{p+1}$ engendré par $\{\mathbf{x}^0, \ldots, \mathbf{x}^p\}$.

Propriétés des estimateurs des MCG

Théorème. (Théorème de Gauss-Markov)

Sous l'hypothèse que le vecteur des résidus $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ est tel que $\mathrm{V}(\epsilon) = \sigma^2\mathbf{\Omega}$ avec $rg(\mathbf{\Omega}) = n$ et $\mathcal{H}_{\mathbf{\Omega}}$, l'estimateur $\mathbf{a}_{mcg} = (\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}Y$ est, parmi les estimateurs linéaires sans biais, celui de variance minimale.

- On obtient par ailleurs, l'estimation MCG $\hat{\sigma}^2_{mcg}$ non biaisée suivante :

$$\hat{\sigma}^2_{mcg} = \frac{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}_{mcg}\|^2_{\mathbf{\Omega}^{-1}}}{n-(p+1)}$$

Lois des estimateurs des MCG

- Si nous faisons de plus l'hypothèse de normalité $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{\Omega})$:

$$\mathbf{a}_{mcg} \sim \mathcal{N}_{p+1}(\mathbf{a}, \sigma^2(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1})$$

$$(n-(p+1))\frac{\sigma^2_{mcg}}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

$\mathbf{a}_{mcg}$ et σ^2_{mcg} sont indépendants

- Les résultats sont similaires à ceux obtenus pour les MCO. Nous pouvons également pratiquer des tests d'hypothèses sur les coefficients de la régression obtenue par MCG à l'aide du résultat général qui suit.

Tests sur les coefficients

- On considère de façon générale un modèle linéaire, $\forall i$:

$$Y_i = X_i^\top\mathbf{a} + \epsilon_i$$

avec les hypothèses $\mathcal{H}_{\mathbf{X}}$ et $\mathcal{H}_\epsilon$ vérifiées.
- Soit une matrice $\mathbf{R}$ de taille $(q \times (p+1))$ et de rang q $(q \leq (p+1))$.
- Soit un vecteur $\mathbf{r}$ de taille q.
- On veut tester :

$$\mathcal{H}_0 : \mathbf{Ra} = \mathbf{r} \text{ contre } \mathcal{H}_1 : \mathbf{Ra} \neq \mathbf{r}$$

- Sous $\mathcal{H}_0$, nous avons la propriété suivante :

$$F = \frac{\|\mathbf{r} - \mathbf{R}\mathbf{a}_{mcg}\|^2_{[\mathbf{R}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}^\top]^{-1}}}{\|\mathbf{y} - \mathbf{X}\mathbf{a}_{mcg}\|^2_{\mathbf{\Omega}^{-1}}} \frac{n-(p+1)}{q} \sim \mathsf{F}_{q,n-(p+1)}$$

- La région de rejet de $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$ au niveau α est :

$$\mathbb{R}_\alpha = \{\hat{F} \geq f_{q,n-(p+1)}(1-\alpha)\}$$

Les MCG en pratique

- Nous avons supposé précédemment que la matrice $\mathbf{\Omega}$ était connue et dans ce cas, nous pouvons en théorie appliquer les MCG et obtenir des estimateurs vérifiant de bonnes propriétés (théorème de Gauss-Markov).
- **Mais en pratique, $\mathbf{\Omega}$ n'est pas connue !**
- Il faut donc une **estimation de la matrice de variance-covariance**.
- Si aucune information n'est supposée sur $\mathbf{\Omega}$, cela implique qu'il faut estimer $n(n+1)/2$ termes correspondant aux éléments $\mathbf{\Omega}_{ij}$ avec $i, j = 1, \ldots, n : i \leq j$ et ceci n'est pas possible en pratique (car plus d'inconnues que d'observations).
- La stratégie est alors de supposer que **$\mathbf{\Omega}$ est représentée par un modèle paramétrique dépendant d'un vecteur de paramètres** θ (avec peu de composantes) qu'il nous faut donc estimer.

Les MCG en pratique (suite)

Nous pouvons considérer deux approches :

1. On suppose la normalité des résidus et on maximise la vraisemblance conjointement par rapport à $\mathbf{a}$ et θ. Il faut dans ce cas utiliser des algorithmes d'optimisation numériques particuliers.

2. On procède en deux étapes et cette approche s'appelle **moindres carrés quasi-généralisés (MCQG)** :
   - On estime d'abord $\hat{\theta}$ et on obtient une estimation $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\theta})$.
   - On remplace ensuite $\hat{\mathbf{\Omega}}$ dans l'estimateur des MCG.

- Dans la suite nous considérons les MCQG.

Les moindres carrés quasi-généralisés (MCQG)

- De façon plus détaillée voici le pseudo-code des MCQG :

Input : $\mathbf{y}, \mathbf{X}, \mathbf{\Omega}(\theta)$ (modèle paramétrique de $\mathbf{\Omega}$)
1 Calcul de $\hat{\mathbf{a}}_{mco} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
2 Calcul de $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}}_{mco}$
3 Calcul de $\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$
4 Calcul de $\hat{\theta}$ à partir de $\hat{\epsilon}$
5 Calcul de $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\theta})$
6 Calcul de $\hat{\mathbf{a}}_{mcqg} = (\mathbf{X}^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{y}$
7 **Output :** $\hat{\mathbf{a}}_{mcqg}$

- L'étape 4 dépend du modèle paramétrique $\mathbf{\Omega}(\theta)$ qui dépend lui même du cas d'étude.

Les moindres carrés quasi-généralisés (MCQG) (suite)

- Plusieurs types de données pouvant nécessiter les MCQG :
  - Cas de données en **coupe transversale** : observations d'un phénomène pour un ensemble d'individu à un instant donné.
    Exemple : on mesure la taille, le poids, d'un ensemble d'individus **à un moment donné**.
  - Cas de données (ou séries) **temporelles** : observations d'un phénomène évoluant dans le temps.
    Exemple : on mesure la concentation en ozone, et la température d'une ville **dans le temps**.
  - Cas de données **géoréférencées** (ou spatiales) : observations d'un phénomène évoluant **dans l'espace**.
    Exemple : on mesure la circonférence, et la hauteur d'un type d'arbres dans plusieurs régions d'un pays.
  - Cas de données en **panel** (ou longitudinales) : observations d'un phénomène pour un ensemble d'individus et **dans le temps**.
    Exemple : on mesure la concentation en ozone, et la température de plusieurs villes dans le temps.
- Nous traiterons les 2 premiers cas et dans des contextes spécifiques.

Les données en coupe transversale avec hétéroscédasticité

- Pour ce type de données, les observations ne sont pas nécessairement distribuées de façon identique et souvent elles correspondent à des individus pouvant appartenir à des groupes distincts et non observés.
- Dans l'exemple précédent, les individus peuvent être des hommes ou des femmes (mais on ne le sait pas) et les distributions de la taille et du poids sont distinctes selon le sexe. Ainsi l'homoscédasticité de la variance résiduelle est une hypothèse vraisemblablement fausse.
- Nous sommes donc dans des cas présentants une **hétéroscédasticité** et pour lesquels il convient d'appliquer les **MCP**.
- Il existe plusieurs modèles paramétriques pour représenter les différentes variances des v.a.r. ϵ_i. Une possibilité est la suivante :

$$\mathrm{V}(\epsilon|\mathbf{X}) = \sigma^2 \exp(b_0 + b_1 \mathbf{x}^1 + \ldots + b_p \mathbf{x}^p) \text{ avec } \theta = (\sigma^2, \mathbf{b})$$

- Dans la suite nous nous intéressons à un cas plus simple mais assez courant.

Les données en coupe transversale avec hétéroscédasticité (suite)

- On considère le cas particulier où les observations sont en fait des agrégats. Typiquement, on suppose que les mesures observées sont des moyennes sur un sous-ensemble d'individus.
- Exemple : on étudie la taille et le poids des individus de l'UE mais les observations que nous avons à disposition correspondent à la taille et au poids moyen dans un pays de l'UE.
- Plus généralement, supposons que nous souhaitons prédire une v.a.r. Y en fonction d'un vecteur aléatoire X selon le modèle linéaire suivant :

$$Y = X^\top \mathbf{a} + \epsilon$$

- En théorie, nous supposons que nous avons un échantillon d'ordre n et nous supposon donc que, $\forall i = 1, \ldots, n$:

$$Y_i = X_i^\top \mathbf{a} + \epsilon_i \text{ où les } \epsilon_i \text{ sont i.i.d. selon une } \mathcal{N}(0, \sigma^2)$$

Les données en coupe transversale avec hétéroscédasticité (suite)

- En revanche, en pratique, nous supposons que nous n'avons pas les réalisations de l'échantillon d'ordre n mais que nous disposons des moyennes selon un partitionnement en q classes de l'échantillon. Soit alors n_j le cardinal de la classe C_j avec $j = 1, \ldots, q$ et $\sum_{j=1}^{q} n_j = n$.
- On a alors comme données un vecteur $\check{\mathbf{y}}$ et une matrice $\check{\mathbf{X}}$ de tailles respectives $(q \times 1)$ et $(q \times q)$ et de termes généraux :

$$\check{y}_j = \frac{1}{n_j} \sum_{i \in C_j} y_i \quad \text{et} \quad \check{x}_{jk} = \frac{1}{n_j} \sum_{i \in C_j} x_{ik}$$

- On doit donc estimer le modèle suivant :

$$\check{\mathbf{y}} = \check{\mathbf{X}}\mathbf{a} + \check{\epsilon} \text{ où } \check{\epsilon} \text{ est de terme général : } \check{\epsilon}_j = \frac{1}{n_j} \sum_{i \in C_j} \epsilon_i$$

Les données en coupe transversale avec hétéroscédasticité (suite)

- Si nous devions estimer la variance résiduelle au sein de chaque classe C_j, nous aurions obtenu une estimation de σ^2 qui dépend de l'effectif de la classe puisque nous avons :

$$\mathrm{V}(\check{\epsilon}_j) = \mathrm{V}(\frac{1}{n_j} \sum_{i \in C_j} \epsilon_i) = \frac{\sigma^2}{n_j}$$

- Le vecteur rédisuel $\check{\epsilon}$ possède alors les propriétés suivantes :
  - $\mathrm{E}(\check{\epsilon}) = 0$.
  - $\mathrm{V}(\check{\epsilon}) = \sigma^2 \mathbf{\Omega}$ avec : $\mathbf{\Omega} = \begin{pmatrix} \frac{1}{n_1} & & \\ & \ddots & \\ & & \frac{1}{n_q} \end{pmatrix}$
- La matrice $\mathbf{\Omega}$ étant connue, on peut donc appliquer les MCP :

$$\hat{\mathbf{a}}_{mcp} = (\check{\mathbf{X}}^\top \mathbf{\Omega}^{-1} \check{\mathbf{X}})^{-1} \check{\mathbf{X}}^\top \mathbf{\Omega}^{-1} \check{\mathbf{y}}$$

Les données avec corrélations temporelles

- Dans ce cas, les observations sont indexés par le temps et nous observons souvent une corrélation temporelle des résidus estimés.
- Il existe dans ce contexte plusieurs modèles paramétriques pour $\mathrm{V}(\epsilon)$. Nous étudions un cas particulier où on suppose que les **résidus suivent un processus autorégressif d'ordre 1 (AR(1))**.
- Formellement on a :

$$Y = X^\top \mathbf{a} + \epsilon$$
$$\forall i = 2, \ldots, n : \epsilon_i = \rho \epsilon_{i-1} + \eta_i$$

avec les hypothèses suivantes concernant le modèle AR(1) :
  - $0 < \rho < 1$ (suite convergente).
  - $\mathrm{V}(\eta) = \sigma^2 \mathbf{I}_n$.
- Dans ce cas, nous avons donc $\theta = (\sigma^2, \rho)$.

Les données avec corrélations temporelles (suite)

- Sous les hypothèses précédentes, on montre que $\mathrm{V}(\epsilon)$ est telle que :

$$\mathbf{\Omega} = \frac{\sigma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ & 1 & \rho & \dots & \rho^{n-2} \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}$$

- La matrice inverse $\mathbf{\Omega}^{-1}$ peut alors s'écrire :

$$\mathbf{\Omega}^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} 1 & -\rho & 0 & \dots & \dots & 0 \\ & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & 0 \\ & & & & 1+\rho^2 & -\rho \\ & & & & & 1 \end{pmatrix}$$

Les données avec corrélations temporelles (suite)

- Nous avons une **solution analytique** de l'inverse de la matrice de variance-covariance du vecteur des erreurs.
- Pour l'instancier, il nous **reste à estimer** ρ. On procède comme décrit dans l'algorithme des MCQG vu précédemment.
- A l'étape 4, on estime ρ par la formule suivante :

$$\hat{\rho} = \frac{\sum_{i=2}^{n} \hat{\epsilon}_i \hat{\epsilon}_{i-1}}{\sum_{i=2}^{n} \hat{\epsilon}_{i-1}^2}$$

- Une fois déterminée $\hat{\rho}$ et $\hat{\mathbf{\Omega}}$ par la suite, nous pouvons déterminer l'estimation des MCQG qui est l'estimation des MCG avec une estimation de $\mathbf{\Omega}$:

$$\hat{\mathbf{a}}_{mcg} = (\mathbf{X}^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{y}$$

Rappel du Sommaire

Introduction

- Nous supposons toujours l'hypothèse de linéarité, $\forall i$:

$$Y_i = X_i^\top \mathbf{a} + \epsilon_i$$

- Mais, nous supposons que l'hypothèse $\mathcal{H}_{\mathbf{X}}$ n'est pas vérifiée.
- Ainsi, $rg(\mathbf{X}) < p$, il existe des dépendances linéaires entre variables explicatives et la matrice $(\mathbf{X}^\top \mathbf{X})^{-1}$ **n'est pas inversible (singulière)**.
- Il n'est donc pas possible de déterminer l'estimateur des MCO.
- Il existe plusieurs types d'extension du modèle linéaire dans ce cas :
  - Les modèles de régression sur composantes principales (PCR).
  - Les modèles de régression "Partial Least Square" (PLS).
  - Les modèles de régression pénalisée (ridge, lasso).
  - ...
- Nous traiterons le cas de la régression sur composantes principales.

$\mathcal{H}_{\mathbf{X}}$ et présence de fortes colinéarités

- S'il existe une dépendance linéaire entre les variables explicatives alors $rg(\mathbf{X}) < p$ (slide précédent).
- Un autre cas où $rg(\mathbf{X}) < p$ est celui pour lequel $n < p$ (plus de variables que d'observations). Il s'agit de problèmes dits de **grandes dimensions**. Dans ce cas, $rg(\mathbf{X}^\top\mathbf{X}) \leq n$ et n'est donc pas inversible.
- Les cas $rg(\mathbf{X}) < p$ qui correspondent à la non satisfaction de $\mathcal{H}_{\mathbf{X}}$ sont des cas extrêmes. Sans aller jusqu'à une dépendance linéaire entre variables explicatives, le cas où il existe de **fortes colinéarités** pose également des problèmes.
- En effet dans ce cas, on montre que $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, la matrice de variance-covariance des coefficients obtenus par les MCO possède des valeurs particulièrement élevées ce qui aboutit à des prédictions moins précises.

$\mathcal{H}_{\mathbf{X}}$ et présence de fortes colinéarités (suite)

- De façon générale, les méthodes que nous verrons sont utilisées dans le cas où on a de fortes colinéarités et/ou des dépendances linéaires entre variables explicatives.
- Pour diagnostiquer ces cas on étudie la **matrice des coefficients de corrélation linéaire** de taille $(p \times p)$, notée $\mathbf{C}$, et de terme générale :

$$\begin{aligned} \mathbf{C}_{ij} &= \rho(X^i, X^j) \\ &= \frac{M_1(X^iX^j) - \overline{X}^i\overline{X}^j}{S(X^i)S(X^j)} \end{aligned}$$

où M_1 est le moment empirique d'ordre r et S est l'écart type empirique (cf slide 56).

- On dira qu'on est en **présence de fortes colinéarités si** :
  - Pour plusieurs paires de variables (X^j, X^k) on observe un fort coefficient de corrélation : $|\rho(X^j, X^k)| \approx 1$.
  - Certaines valeurs propres de la matrice des coefficients de corrélation sont très proches de 0.

Introduction (suite)

- La régression PCR corrige des défauts intrinsèques aux données mais, en contre-partie, **le théorème de Gauss-Markov n'est plus vérifié** : l'estimateurs PCR est biaisé contrairement aux MCO.
- De plus, l'estimateur PCR n'est pas invariante à un changement d'échelle. Ainsi, il est d'usage de travailler sur la **matrice de données centrées réduites**. Pour ne pas allourdir les formules nous utiliserons les mêmes notations que précédemment.
  Donc, $\mathbf{y}$ et $\mathbf{X}$ sont les données centrées et réduites.
- Plus formellement, soit $\mathbf{v}$ de taille $(n \times 1)$ les observations de la variable à expliquer initiale et $\mathbf{U}$ de taille $(n \times p)$, les observations des variables explicatives initiales. Les termes généraux de $\mathbf{y}$ et $\mathbf{X}$ sont :

$$\forall i : y_i = \frac{v_i - \overline{\mathbf{v}}}{\sigma_{\mathbf{v}}} \quad \text{et} \quad \forall i,j : x_{ij} = \frac{u_{ij} - \overline{u}^j}{\sigma_{\mathbf{u}^j}}$$

où $\overline{\mathbf{v}}$ et $\sigma_{\mathbf{v}}$ sont les moyennes et écart-types de $\mathbf{v}$.

Régression PCR

- Le modèle linéaire classique s'écrit :

$$Y = a_1X^1 + \ldots + a_pX^p + \epsilon$$

- En raison des problèmes de singularité, on cherche donc un modèle :

$$Y = a_1\check{X}^1 + \ldots + a_p\check{X}^p + \epsilon$$

où les $\check{X}^j$ sont de nouvelles variables qui sont non colinéaires.

- $\mathbf{X}^\top\mathbf{X}$ est symmétrique et réelle, elle est donc **diagonalisable**[5] :

$$\mathbf{X}^\top\mathbf{X} = \mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}^{-1}$$

où $\mathbf{D}$ est la matrice diagonale remplie des valeurs propres et $\mathbf{\Lambda}$ de taille $(p \times p)$ est la matrice des vecteurs propres normés.

- La matrice $\mathbf{\Lambda}$ est orthogonale et donc :

$$\mathbf{\Lambda}\mathbf{\Lambda}^\top = \mathbf{\Lambda}^\top\mathbf{\Lambda} = \mathbf{I}_p \text{ et } \mathbf{\Lambda}^\top = \mathbf{\Lambda}^{-1}$$

5. Ces propriétés sont similaires à celles de la matrice de variance-covariance des résidus $\mathbf{\Omega}$ étudiées dans le cas des MCG.

Régression PCR (suite)

- La décomposition précédente est similaire à une **ACP** (**Analyse en Composantes Principales**). La matrice $\mathbf{X}^\top\mathbf{X}$ **est la matrice de corrélation** des variables et dans ce cas, **les composantes principales sont obtenues à l'aide des vecteurs propres**.
- Puisque $\mathbf{\Lambda}\mathbf{\Lambda}^\top = \mathbf{I}_p$, nous avons l'identité suivante :

$$Y = \mathbf{X}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{a} + \epsilon$$

- Posons alors :

$$\check{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda} \text{ et } \check{\mathbf{a}} = \mathbf{\Lambda}^\top\mathbf{a}$$

- L'équation précédente s'écrit donc :

$$Y = \check{\mathbf{X}}\check{\mathbf{a}} + \epsilon$$

- Ce modèle suggère une **régression par les MCO sur les composantes principales** car $\check{\mathbf{X}}$ représente les coorodonnées des individus sur les axes principaux issus de l'ACP.

Régression PCR (suite)

- Par construction, nous avons :

$$\begin{aligned}\check{\mathbf{X}}^\top\check{\mathbf{X}} &= (\mathbf{X}\mathbf{\Lambda})^\top(\mathbf{X}\mathbf{\Lambda}) \\ &= \mathbf{\Lambda}^\top\mathbf{X}^\top\mathbf{X}\mathbf{\Lambda} \\ &= \mathbf{\Lambda}^\top\mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}^\top\mathbf{\Lambda} \\ &= \mathbf{D}\end{aligned}$$

où

$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^p \end{array}\begin{array}{c} \begin{array}{ccc} \mathbf{v}^1 & \dots & \mathbf{v}^p \end{array} \\ \begin{pmatrix} v_{11} & \dots & v_{1p} \\ \vdots & \dots & \vdots \\ v_{n1} & \dots & v_{np} \end{pmatrix} \end{array} ; \mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \text{ et } \lambda_1 \geq \ldots \geq \lambda_p.$$

- Les colonnes de $\check{\mathbf{X}} = \begin{pmatrix}\check{\mathbf{x}}^1 & \dots & \check{\mathbf{x}}^p\end{pmatrix}$ forment une base de $\mathbb{R}^p$ qui est **orthogonale** (donc les nouvelles variables $\check{\mathbf{x}}^j$ sont **mutuellement indépendantes** contrairement aux variables initiales $\mathbf{x}^j$).
- Les normes des vecteurs $\check{\mathbf{x}}^j$ sont respectivement les valeurs propres λ_j.

Régression PCR (suite)

- Comme en ACP, la régression PCR a vocation à ne pas utiliser toutes les composantes principales.
- **On retient** $k < p$ **composantes principales**. Celles-ci rassemblent une **partie de l'information** contenue au sein des variables explicatives.
- De ce fait, les $(p-k)$ composantes non sélectionnées, représente la part de l'information que l'on décide de ne pas tenir compte. On considère cette partie comme non pertinente pour la modélisation.
- Le concept information utilisée ici est relatif à la **variance** (inertie ou dispersion) du nuage des observations.
- On a la propriété que la variance associée à une composante principale est donnée par la valeur propre qui lui est associée.
- Plus la valeur de λ_j est forte, plus l'information apportée par la composante (en terme de variance) est importante.
- Naturellement, nous retenons les composantes les plus informatives et donc les k vecteurs associées aux k plus grandes valeurs propres.

Régression PCR (suite)

- Notons la matrice des k premières composantes principales par :

$$\tilde{\mathbf{X}} = \begin{pmatrix}\check{\mathbf{x}}^1 & \dots & \check{\mathbf{x}}^k\end{pmatrix} = \mathbf{X}\check{\mathbf{\Lambda}}$$

où $\check{\mathbf{\Lambda}} = \begin{pmatrix}\mathbf{v}^1 & \dots & \mathbf{v}^k\end{pmatrix}$ (k premiers vecteurs propres de $\mathbf{\Lambda}$).

- Le modèle s'écrit alors :

$$Y = \tilde{\mathbf{X}}\tilde{\mathbf{a}} + \epsilon$$

où $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_k)$ est le vecteur des coefficients d'ordre $k < p$.

- On applique alors les MCO dans ce sous-espace à k dimensions :

$$\tilde{\mathbf{a}}_{mco} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top Y$$

- Les variables $\check{\mathbf{x}}^j$ étant orthogonales deux à deux, on a la propriété suivante :

$$\mathrm{C}(\tilde{a}_{i,mco}, \tilde{a}_{j,mco}) = \begin{cases} \frac{\sigma^2}{\lambda_i} \text{ si } i = j \\ 0 \text{ sinon} \end{cases}$$

Régression PCR (suite)

- Nous pouvons revenir aux variables initiales (centrées-réduites), puisque les composantes principales sont des combinaisons linéaires des variables explicatives initiales !
- Rappelons que $\check{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda}$. Les colonnes de $\mathbf{\Lambda} = \begin{pmatrix} \mathbf{v}^1 & \ldots & \mathbf{v}^p \end{pmatrix}$ sont les vecteurs propres de $\mathbf{X}^\top\mathbf{X}$. Par ailleurs, comme $\check{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda}$, nous avons $\forall j \ : \check{\mathbf{x}}^j = \mathbf{X}\mathbf{v}^j$ qui sont les composantes principales.
- Rappelons aussi que $\tilde{\mathbf{X}} = \mathbf{X}\check{\mathbf{\Lambda}} = \begin{pmatrix} \check{\mathbf{x}}^1 & \ldots & \check{\mathbf{x}}^k \end{pmatrix}$.
- Par conséquent, on peut revenir à l'espace initial comme suit :

$$\begin{aligned} Y &= \tilde{\mathbf{X}}\tilde{\mathbf{a}}_{mco} + \epsilon \\ &= \mathbf{X}\check{\mathbf{\Lambda}}\tilde{\mathbf{a}}_{mco} + \epsilon \\ &= \mathbf{X}\mathbf{a}_{pcr} + \epsilon \end{aligned}$$

Régression PCR (suite)

- L'estimateur de la régression PCR, $\mathbf{a}_{pcr}$, est le vecteur des coefficients estimé par les MCO dans l'espace de dimension k engendré par les vecteurs propres de $\mathbf{X}^\top\mathbf{X}$ (ie $\tilde{\mathbf{a}}_{mco}$) qui est exprimé dans l'espace de dimension p engendré par les variables initiales (ie $\check{\mathbf{\Lambda}}\tilde{\mathbf{a}}_{mco}$) :

$$\mathbf{a}_{pcr} = \check{\mathbf{\Lambda}}\tilde{\mathbf{a}}_{mco} = \check{\mathbf{\Lambda}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top Y$$

- L'estimation est obtenue en remplaçant Y par la variable observée $\mathbf{y}$:

$$\hat{\mathbf{a}}_{pcr} = \check{\mathbf{\Lambda}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y}$$

- En ce qui concerne la prédiction pour les observations, nous avons :

$$\hat{\mathbf{y}} = \check{\mathbf{X}}\hat{\tilde{\mathbf{a}}}_{mco} = \mathbf{X}\hat{\mathbf{a}}_{pcr}$$

- Si nous souhaitons revenir à l'échelle initiale, nous avons alors la formule suivante :

$$\hat{\mathbf{y}} = \bar{\mathbf{y}}\mathbf{1}_n + \hat{\sigma}_{\mathbf{y}}\mathbf{X}\hat{\mathbf{a}}_{pcr}$$

Quelques références

- Guyon, X., 2001. *Statistique et économétrie*, Ellipses
- Cornillon, P-A. et Matzner-Lober, E., 2011. *Régression avec R*, Springer
- Saporta, G., 2006. *Probabilites, Analyse des Données et Statistique*, Technip
- Seber, G.A.F et Lee, A.J. 2003. *Linear Regression Analysis*, Wiley