Modélisation Statistique et Analyse de Données

Auteurs : Jean-Baptiste Durand – Ollivier Taramasco – Olivier Gaudoin

Contributeur : Christophe Dutang

ENSIMAG - 2ème année

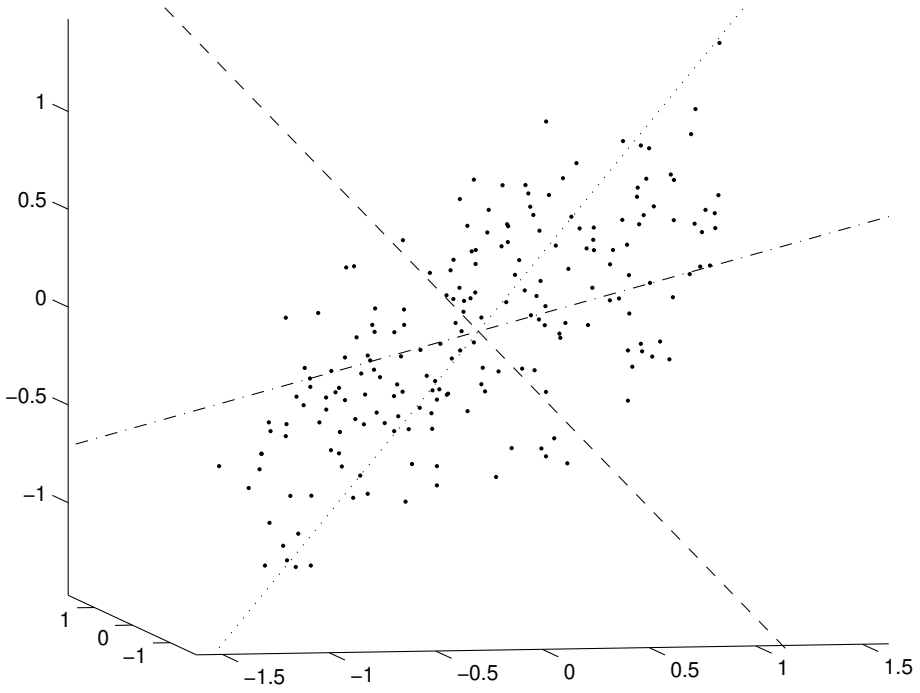

Table des matières

Introduction générale

Probabilités

— calcul des probabilités : mathématiques pures, déconnecté du concret.
— probabilités appliquées : modèles probabilistes pour phénomènes concrets et aléatoires. On définie les variables aléatoires dont on connaît les lois de probabilité et on effectue divers calculs dessus.
— processus aléatoires : modéliser des variables aléatoires qui évoluent dans le temps ou dans l'espace. Cela permet, préalablement à toute expérience, de faire des prévisions sur le résultat d'une expérience aléatoire.

Statistique

La statistique part de données : résultats d'expériences, enquêtes socio-économique, séries financières *etc.* Un des objectifs de la statistique est d'extraire le maximum d'informations de ces données.

Types de données :
— données unidimensionnelles : $x_1, x_2, \ldots, x_n \in \mathbb{R}$
— données bidimensionnelles : $(x_1, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^2$
— données multidimensionnelles : $(x_i^j) \in \mathcal{M}_{n,p}(\mathbb{R})$

Les observations sont considérées comme des réalisations de variables aléatoires.

Classes de méthodes statistiques

Statistique inférentielle

Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations.

Statistique paramétrique

On suppose que l'on connaît la famille de lois de probabilités à laquelle appartient la loi des variables aléatoires dont les données sont des réalisations. La seule inconnue est la valeur du paramètre θ de cette loi.

Principaux problèmes :
— estimation de θ
— tests d'hypothèse

Avec des données multidimensionnelles, il faut faire attention à la dépendance des variables.

Statistique non paramétrique

On utilise la statistique non paramétrique lorsqu'on ne connaît pas parfaitement la famille de lois des probabilités des observations.
En statistique non paramétrique, il y a également des problèmes d'estimation et de test, mais ce qu'on estime n'est pas un paramètre. On peut choisir un modèle probabiliste adapté à ces données (test d'adéquation), estimer la densité des observations (inférence fonctionnelle), comparer des échantillons, *etc.*

Statistique descriptive ou Analyse des données

L'Analyse des données a pour objectif de caractériser l'information contenue dans un ensemble de variables et pour lesquelles on ne dispose pas, en général, de modèles. Elle permet, par exemple, de regrouper des variables en des facteurs communs, ou d'identifier les relations de dépendance entre ces variables.

Outils : représentations graphiques, projections, géométrie, *etc.*

Analyse Statistique Multidimensionnelle

C'est l'étude de la statistique sur données à plusieurs dimensions sous les deux aspects, inférentiels (modèles linéaires) et descriptifs.

Vocabulaire

En statistique multidimensionnelle, les données sont constituées de *tableaux de mesures* faites sur n *individus* (ou *unités statistiques*). Sur chacun de ces n individus, on mesure p *variables*. L'ensemble des individus constitue *l'échantillon*.

L'échantillon est donc un sous-ensemble d'un ensemble plus vaste appelé *population (mère)*. La population peut être finie même si le nombre d'individus qui la composent, est très grand (par exemple, la population française). Elle peut aussi être infinie : la population des prix d'un actif à la Bourse est *a priori* l'ensemble des nombres réels positifs.

Lorsque l'échantillon est constitué de la totalité des individus de la population, on dit que l'on fait *un recensement*. C'est un cas très rare, à la fois pour des raisons de faisabilité et surtout de coût.

Lorsque l'échantillon n'est pas la population mère, on dit qu'on a fait un *sondage*. Le but de ce sondage est de constituer un *échantillon représentatif* de la population de façon à pouvoir étendre les enseignements obtenus à l'ensemble de cette population. Il existe des méthodes pour obtenir des échantillons représentatifs qui ne seront pas traitées dans ce cours. Dans l'exemple 5, il est peu probable que les pays présents dans l'échantillon soient représentatifs de l'ensemble des pays du monde. Par contre, les étudiants interrogés dans l'exemple 6 pourraient être représentatifs de l'ensemble de la population étudiante française.

Les types de variables

La distinction entre variables qualitatives et quantitatives est fondamentale : on ne traite pas de la même manière ces deux types de variables. Pour simplifier, les variables quantitatives sont celles sur

lesquelles on peut faire des calculs arithmétiques et pour lesquelles, ces calculs ont un sens. Toutes les autres variables sont qualitatives ([1]).

Les variables qualitatives peuvent être séparées en deux sous-types :

- Les variables nominales : elles ne peuvent pas être ordonnées (par exemple, les couleurs ou les variables booléennes muettes)
- Les variables ordinales : elles peuvent être ordonnées mais aucune opération arithmétique n'a de sens sur ces variables (par exemple, une variable à valeurs dans l'ensemble {petit, moyen, grand}).

Les variables quantitatives peuvent être elles aussi séparées en deux sous-types :

- Les échelles d'intervalles : ces variables possèdent une unité mais pas de "zéro" naturel (par exemple, l'échelle de température).
- Les échelles de proportion : ce sont les variables quantitatives qui possèdent un "zéro" naturel (par exemple, la taille en cm).

Le tableau suivant montre quelles statistiques descriptives il est possible de calculer à partir de l'observation d'un échantillon des divers types de variables.

Type d'échelle	Propriétés	Statistiques descriptives
Nominale	Relation d'équivalence	Mode, fréquence
Ordinale	Relation d'ordre	Médiane, fractiles
Intervalle	Pas de zéros	Moyenne, variance
Proportion	Toute opération	Moyenne, variance

TABLE 1 – niveau de mesures des variables

Remarque 1. *Le même mot "variable" désigne d'une part la grandeur que l'on veut étudier (variable observée ou variable statistique) et l'objet mathématique qui le modélise (variable aléatoire). Le contexte permet d'éviter la confusion entre ces deux notions.*

1. Il n'est pas rare que des variables qualitatives soient **codées** à l'aide de nombres. Il est alors possible d'effectuer n'importe quelle opération sur ces nombres, mais le résultat a, en général, aucun sens.

Chapitre 1

Régression linéaire multiple

1.1 Rappels sur les vecteurs aléatoires

Un vecteur aléatoire Z peut être caractérisé par

— sa fonction de densité

$$f_Z(z) = f_{Z_1,\dots,Z_d}(z_1,\dots,z_d)$$

— sa fonction de masse de probabilité

$$p_Z(z) = P(Z_1 = z_1,\dots,Z_d = z_d).$$

Deux lois usuelles sont la loi normal multivariée et la loi multinomiale.
Le vecteur d'espérance et la matrice de covariance sont définis par

$$\mathrm{E}(Z) = \begin{pmatrix} \mathrm{E}[Z_1] \\ \vdots \\ \mathrm{E}[Z_d] \end{pmatrix}, K_Z = \begin{pmatrix} Cov(Z_1,Z_1) & \dots & Cov(Z_1,Z_d) \\ & \vdots & \\ Cov(Z_d,Z_1) & \dots & Cov(Z_d,Z_d) \end{pmatrix} = \mathrm{E}[(Z-\mathrm{E}(Z))(Z-\mathrm{E}(Z))^T].$$

On a les propriétés suivantes

— Si A est une matrice déterministe et Z un vecteur aléatoire de matrice de covariance K_Z alors le vecteur $A\,Z$ a pour espérance $\mathrm{E}(A\,Z) = A\,\mathrm{E}(Z)$ et pour matrice de covariance $K_{A\,Z} = A\,K_Z\,{}^tA$.

Définition 1. *Un vecteur gaussien aléatoire Z est dit gaussien si toute combinaison linéaire de ses composantes suit une loi normale.*

Proposition 1. *Les composantes d'un vecteur gaussien Z sont indépendantes si et seulement si la matrice de covariance K_Z est diagonale.*

Proposition 2. *Un vecteur gaussien d'espérance μ et de matrice de covariance Σ admet pour densité*

$$f_Z(z) = \frac{1}{(2\pi)^{d/2}\sqrt{|det(\Sigma)|}}\exp(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)).$$

Proposition 3. *Si $Z \sim \mathcal{N}_d(\mu,\Sigma)$ alors*

$$(Z-\mu)^T\Sigma^{-1}(Z-\mu) \sim \chi^2_d.$$

Théorème 1 (de Cochran)**.** *Considérons un vecteur gaussien $Z \sim \mathcal{N}(\mu,\sigma^2 I_d)$ et une matrice P de projection orthogonale sur un sous-espace vectoriel de $\mathbb{R}^d$ de dimension p. Pour $P_\perp = I_d - P$, on a*

1. *$PZ \sim \mathcal{N}(P\mu,\sigma^2 P)$ et $P_\perp Z \sim \mathcal{N}(P_\perp\mu,\sigma^2 P_\perp)$.*
2. *les vecteurs PZ et $P_\perp Z$ sont indépendants.*
3. *les distances $\|P(Y-\mu)\|^2/\sigma^2 \sim \chi_p$ et $\|P_\perp(Y-\mu)\|^2/\sigma^2 \sim \chi_{d-p}$.*

1.2 Modèle de régression linéaire multiple

Dans le modèle de régression linéaire simple, la variable à expliquer Y ne dépendait que d'un seul facteur x. Dans le modèle de régression multiple, elle dépend de $p \geq 1$ prédicteurs $x_1, \ldots, x_p$.

La liaison entre la variable explicative et les prédicteurs est encore affine :

$$\mathcal{Y} = \beta_p\, x_p + \beta_{p-1}\, x_{p-1} + \cdots + \beta_1\, x_1 + \beta_0 + \varepsilon$$

Si on dispose de plusieurs réalisations, on obtient le modèle suivant :

Définition 2. *Le modèle linéaire de régression multiple (ou modèle linéaire général) est défini par :*

$$\mathcal{Y}_i = \beta_p\, x_{p,i} + \beta_{p-1}\, x_{(p-1),i} + \cdots + \beta_1\, x_{1,i} + \beta_0 + \varepsilon_i,\ \ i = 1, \ldots, n$$

où les résidus ε_i sont des variables aléatoires centrées.

Contrairement au modèle linéaire simple, les résidus ne sont pas supposés être indépendants et de même loi.

Définition 3 (Écriture matricielle du modèle linéaire général)**.**

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

— $\boldsymbol{Y} = \begin{pmatrix} \mathcal{Y}_1 \\ \vdots \\ \mathcal{Y}_n \end{pmatrix}$ *est un vecteur aléatoire des observations dans* $\mathbb{R}^n$.

— $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ *est le vecteur des paramètres inconnus dans* $\mathbb{R}^{p+1}$.

— $$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} \end{pmatrix}$$

$\boldsymbol{X}$ est une matrice déterministe connue d'ordre $n\times(p+1)$ ($\mathbb{R}^{n\times p}$) appelée matrice des prédicteurs ou matrice des régresseurs ou matrice du plan d'expérience ou matrice des variables explicatives.

— $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ *est un vecteur aléatoire de $\mathbb{R}^n$, centré, appelé vecteurs des résidus.*

Le vecteur des observations $\boldsymbol{Y}$ a pour espérance $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\,\boldsymbol{\beta}$ et pour matrice de covariance $K_Y = K_{\boldsymbol{\varepsilon}}$ (*i.e.* la matrice de covariance du vecteur $\boldsymbol{\varepsilon}$). L'étude du modèle linéaire va dépendre fortement des hypothèses effectuées sur cette matrice.

On supposera toujours que la matrice des régresseurs est de range $p+1$ de telle sorte que $X^T X \in \mathbb{R}^{p\times p}$ soit inversible.

Le cas le plus simple est celui qui généralise le modèle linéaire simple : les résidus ε_i sont indépendants de même loi, leur variance est notée σ^2. La matrice de covariance du vecteur $\boldsymbol{\varepsilon}$ est alors $K_{\boldsymbol{\varepsilon}} = \sigma^2\, I_n$

où I_n est la matrice unité d'ordre n. Dans ce cas, la méthode utilisée est celle des *moindres carrés ordinaires (*M.C.O.*)*.

Il est aussi possible d'envisager le cas où les résidus ne sont ni indépendants ni de même loi. Dans ces conditions, la matrice de covariance du vecteur $\boldsymbol{\varepsilon}$ est une matrice symétrique semi-définie positive d'ordre n, quelconque. Cette matrice est généralement notée $K_{\boldsymbol{\varepsilon}} = V$ et la méthode est celle des *moindres carrés généralisés (*M.C.G.*)*.

Il est possible de traiter le cas où les prédicteurs sont des variables aléatoires. Comme dans le cas à deux dimensions, on montre que la meilleure prévision $\widehat{\mathcal{Y}}$ de $\mathcal{Y}$ connaissant $\mathcal{X}_1, \ldots, \mathcal{X}_p$ est $\mathrm{E}\left[\mathcal{Y}|\mathcal{X}_1, \ldots, \mathcal{X}_p\right]$. De même, si le vecteur $(\mathcal{X}_1, \ldots, \mathcal{X}_p, \mathcal{Y})$ est un vecteur gaussien, on montre que $\widehat{\mathcal{Y}}$ est une fonction affine de $(\mathcal{X}_1, \ldots, \mathcal{X}_p)$.

Les régresseurs peuvent être indépendants ou pas. Par exemple, le *modèle de régression polynômiale* consiste à choisir $x_j = x^j, \quad \forall j \in \{1 \ldots p\}$. Autrement dit, le problème est de choisir le «meilleur» polynôme de degré p qui passe à travers le nuage $\{(x_i, y_i)\,, i = 1 \ldots n\}$ de $\mathbb{R}^2$.

1.3 Moindres carrés ordinaires (M.C.O.)

Dans ce paragraphe, la matrice de covariance des résidus est $K_{\boldsymbol{\varepsilon}} = \sigma^2\, I_n$. Le problème qui se pose est l'estimation des $p+2$ paramètres réels $\beta_0, \ldots, \beta_p, \sigma^2$ ou encore du vecteur $\boldsymbol{\beta}$ de $\mathbb{R}^{p+1}$ et du réel σ^2.

1.3.1 Estimateurs des moindres carrés ordinaires

La méthode des moindres carrés vue dans le cas du modèle linéaire simple va se généraliser sans problème dans le cas de la régression multiple. Il faut trouver le vecteur $\widehat{\boldsymbol{B}} \in \mathbb{R}^{p+1}$ tel que $\boldsymbol{X}\,\widehat{\boldsymbol{B}}$ soit le plus proche possible de $\boldsymbol{Y}$ au sens de la norme euclidienne de $\mathbb{R}^n$ parmi tous les vecteurs de $\mathbb{R}^n$ qui s'écrivent $\boldsymbol{X}\,\boldsymbol{\beta}$.

Définition 4. *L'estimateur des* M.C.O. *de $\boldsymbol{\beta}$ est le vecteur $\widehat{\boldsymbol{B}} \in \mathbb{R}^{p+1}$ défini par :*

$$\widehat{\boldsymbol{B}} = \mathrm{Arg} \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} ||\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\beta}||^2$$

Soit $E_X = \{\boldsymbol{X}\boldsymbol{\beta}\,;\, \boldsymbol{\beta} \in \mathbb{R}^{p+1}\} \subset \mathbb{R}^n$ le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\boldsymbol{X}$. Alors $\boldsymbol{X}\,\widehat{\boldsymbol{B}}$ est la projection orthogonale de $\boldsymbol{Y}$ sur E_X.

Si les colonnes de $\boldsymbol{X}$ sont linéairement indépendantes, alors $\dim E_X = p+1$ (sinon on a un *problème de multicolinéarité*). On supposera dans la suite que $\dim E_X = p+1$.

Proposition 4 (Équation normale)**.** *$\widehat{\boldsymbol{B}}$ est une solution de l'équation :*

$${}^t\boldsymbol{X}\,\boldsymbol{X}\,\boldsymbol{\beta} = {}^t\boldsymbol{X}\,\boldsymbol{Y}$$

appelée équation normale du modèle linéaire.

Théorème 2 (de Gauss-Markov)**.** *Si ${}^t\boldsymbol{X}\,\boldsymbol{X}$ est inversible, $\widehat{\boldsymbol{B}} = \left({}^t\boldsymbol{X}\,\boldsymbol{X}\right)^{-1} {}^t\boldsymbol{X}\,\boldsymbol{Y}$ est appelé estimateur de Gauss-Markov. $\widehat{\boldsymbol{B}}$ est un estimateur sans biais de $\boldsymbol{\beta}$, de variance minimum parmi tous les estimateurs sans biais linéaires et sa matrice de covariance est*

$$K_{\widehat{\boldsymbol{B}}} = \sigma^2 \left({}^t\boldsymbol{X}\,\boldsymbol{X}\right)^{-1}$$

Pour estimer $\sigma^2 = \mathrm{Var}(\varepsilon_i)$, on se propose, comme dans le cas de la régression simple, d'utiliser la variance empirique des résidus empiriques.

Définition 5. *On appelle résidus empiriques les variables aléatoires*

$$\widehat{\varepsilon}_i = \mathcal{Y}_i - \widehat{\mathcal{Y}}_i = \mathcal{Y}_i - \left(\boldsymbol{X}\,\widehat{\boldsymbol{B}}\right)_i = \mathcal{Y}_i - \widehat{\boldsymbol{B}}_p\, x_{p\,i} - \ldots - \widehat{\boldsymbol{B}}_1\, x_{1\,i} - \widehat{\boldsymbol{B}}_0$$

La variance résiduelle $S^2_{Y|x}$ est la variance empirique des $\widehat{\varepsilon}_i$.

Proposition 5.

$$\forall i \in \mathbb{R}^n,\ \mathrm{E}\left[\widehat{\varepsilon}_i\right] = 0$$

$$Si\ \mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in E_X,\ alors \begin{cases} \overline{\widehat{\varepsilon}}_n = \frac{1}{n}\sum\limits_{i=1}^{n} \widehat{\varepsilon}_i = 0 \\ S^2_{Y|x} = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{B}}\|^2 \end{cases}$$

Proposition 6. $\widehat{\sigma}^2 = \frac{n}{n-p-1}\, S^2_{Y|x} = \frac{1}{n-p-1}\, ||\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{B}}||^2$ *est un estimateur sans biais de σ^2.*

Cela suppose implicitement que $n > p+1$. Ce résultat n'est pas surprenant puisqu'on ne peut espérer estimer plus de paramètres qu'il n'y a d'observations.

1.3.2 Cas du modèle linéaire simple

Le cas du modèle linéaire simple $\mathcal{Y}_i = \beta_1\, x_i + \beta_0 + \varepsilon_i,\ i = 1, \ldots, n$ correspond au cas $p = 1$. L'écriture matricielle de ce modèle est donc $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec :

$$\boldsymbol{Y} = \begin{pmatrix} \mathcal{Y}_1 \\ \vdots \\ \mathcal{Y}_n \end{pmatrix},\ \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

1.4 Modèle linéaire gaussien

1.4.1 Définition du modèle et estimation des paramètres

Définition 6. *Le modèle de régression linéaire multiple gaussien (ou modèle linéaire gaussien) est défini par :*

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où $\boldsymbol{\varepsilon}$ est un vecteur gaussien centré dans $\mathbb{R}^n$ de matrice de covariance $K_{\boldsymbol{\varepsilon}}$.

Dans la suite, nous nous contenterons d'étudier le cas où $K_{\boldsymbol{\varepsilon}} = \sigma^2\, I_n$, ce qui signifie que les résidus ε_i sont mutuellement indépendants et de même loi $\mathcal{N}(0,\, \sigma^2)$.

Remarque 2.

$$\boldsymbol{Y} \rightsquigarrow \mathcal{N}_{\mathbb{R}^n}(\boldsymbol{X}\,\boldsymbol{\beta},\, \sigma^2\, I_n)$$

Proposition 7. *Les estimateurs de maximum de vraisemblance de $\boldsymbol{\beta}$ et σ^2 sont respectivement $\widehat{\boldsymbol{B}}$ et $\frac{n-p-1}{n}\,\widehat{\sigma}^2$ (qui est biaisé).*

Proposition 8.

$$\widehat{\boldsymbol{B}} = ({}^t\boldsymbol{X}\,\boldsymbol{X})^{-1}\,{}^t\boldsymbol{X}\,\boldsymbol{Y} \rightsquigarrow \mathcal{N}_{\mathbb{R}^{p+1}}\left(\boldsymbol{\beta},\, \sigma^2\,({}^t\boldsymbol{X}\,\boldsymbol{X})^{-1}\right)$$

En particulier, chaque composante $\hat{B}_i$ de $\widehat{\boldsymbol{B}}$ est un estimateur sans biais de β_i, de loi normale et de variance calculable. Cela permet de construire des intervalles de confiance et des tests d'hypothèse sur ces paramètres.

Proposition 9. *$\widehat{\boldsymbol{B}}$ est l'estimateur sans biais de variance minimale de $\boldsymbol{\beta}$ parmi tous les estimateurs sans biais de $\boldsymbol{\beta}$ (et pas seulement parmi les estimateurs sans biais linéaires).*

Proposition 10. *$\boldsymbol{X}\widehat{\boldsymbol{B}}$ et $\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ sont indépendants.*

Corollaire 1. *$\widehat{\boldsymbol{B}}$ et $\widehat{\sigma}^2$ sont indépendants*

Proposition 11.

$$\frac{n-p-1}{\sigma^2}\widehat{\sigma}^2 = \frac{1}{\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{B}}\|^2 \rightsquigarrow \mathcal{X}^2_{n-p-1}$$

Ce résultat permet d'obtenir facilement des intervalles de confiance pour σ^2.

1.4.2 Décomposition de la variance et test de pertinence de la régression

Soient $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\boldsymbol{B}}$ et $\overline{\mathcal{Y}} = \begin{pmatrix} \overline{\mathcal{Y}}_n \\ \vdots \\ \overline{\mathcal{Y}}_n \end{pmatrix} = \overline{\mathcal{Y}}_n\,\mathbb{1}_n$. On a $\overline{\mathcal{Y}} \in E_X$ (sous l'hypothèse que $\mathbb{1}_n \in E_X$).

La *variance totale* est la variance empirique de la variable à expliquer $\boldsymbol{Y}$:

$$S^2_Y = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{Y}_i - \overline{\mathcal{Y}}_n\right)^2 = \frac{1}{n}\|\boldsymbol{Y} - \overline{\mathcal{Y}}\|^2$$

La *variance expliquée* est la variance empirique des valeurs prédites $\widehat{Y}_i$:

$$S^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\mathcal{Y}}_n\right)^2 = \frac{1}{n}\|\widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}}\|^2$$

La *variance résiduelle* est la variance empirique des résidus empiriques $\widehat{\varepsilon}_i = \mathcal{Y}_i - \widehat{Y}_i$:

$$S^2_{Y|x} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{Y}_i - \widehat{Y}_i\right)^2 = \frac{1}{n}\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$$

Proposition 12 (Formule de décomposition de la variance)**.**

$$\|\boldsymbol{Y} - \overline{\mathcal{Y}}\|^2 = \|\widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}}\|^2 + \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$$

$$\textit{ou encore : } S_Y^2 = S_{\widehat{Y}}^2 + S_{Y|x}^2$$

Proposition 13. *Si* $\beta_p = \ldots = \beta_1 = 0$ *alors*

$$\frac{1}{\sigma^2}\left\|\boldsymbol{Y} - \overline{\mathcal{Y}}\right\|^2 \rightsquigarrow \mathcal{X}^2_{n-1} \textit{ et } \frac{1}{\sigma^2}\left\|\widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}}\right\|^2 \rightsquigarrow \mathcal{X}^2_p$$

Corollaire 2. *Sous* $\mathrm{H}_0 : \beta_p = \ldots = \beta_1 = 0$

$$\frac{n-p-1}{p}\frac{\left\|\widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}}\right\|^2}{\left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2} = \frac{n-p-1}{p}\frac{S_{\widehat{Y}}^2}{S_{Y|x}^2} \rightsquigarrow \mathcal{F}(p, n-p-1)$$

Cette propriété permet de construire le test $\mathrm{H}_0 : \beta_p = \ldots = \beta_1 = 0$ contre $\mathrm{H}_1 : \overline{\mathrm{H}_0}$ qui est équivalent au test de *pertinence de la régression* dans le modèle linéaire simple.

Ce dernier test pouvait s'exprimer à l'aide du coefficient de corrélation linéaire empirique, car pour $p = 1$, $(n-2)\,\frac{S_{\widehat{Y}}^2}{S_{Y|x}^2} = (n-2)\,\frac{R_{xY}^2}{1-R_{xY}^2}$ et $R_{xY}^2 = \frac{S_{\widehat{Y}}^2}{S_Y^2}$.

Nous allons définir une notion analogue dans le cas de la régression linéaire multiple.

Définition 7. *On appelle coefficient de corrélation linéaire multiple entre la variable à expliquer* $\boldsymbol{Y}$ *et les régresseurs* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ *la variable aléatoire :*

$$R = \sup_{(a_1,\ldots,a_p)\in\mathbb{R}^p} R_{\boldsymbol{Y}, \sum_{j=1}^{p} a_j \boldsymbol{X}_j}$$

R est donc la valeur maximale prise par le coefficient de corrélation linéaire empirique entre la variable $\boldsymbol{Y}$ et n'importe quelle combinaison linéaire des $\boldsymbol{X}_j$.

Il est facile de montrer que :
- R est à valeur dans $[0,1]$
- $R = 1 \Leftrightarrow \exists (a_0, a_1, \ldots, a_p) \in \mathbb{R}^{p+1}, \boldsymbol{Y} = a_0 + \sum_{j=1}^{p} a_j \boldsymbol{X}_j$. Autrement dit, $R = 1$ s'il n'y a pas de bruit.

On construit un test de pertinence de la régression en admettant que la régression est pertinente quand R est significativement proche de 1.

On peut montrer que :
- $R = \cos(\boldsymbol{Y} - \overline{\mathcal{Y}}, \widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}})$
- $R^2 = \frac{S^2_{\widehat{Y}}}{S^2_Y}$
- $\frac{R^2}{1-R^2} = \frac{S^2_{\widehat{Y}}}{S^2_{Y|x}} = \frac{\|\widehat{\boldsymbol{Y}} - \overline{\mathcal{Y}}\|^2}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2}$

Le test de la pertinence de la régression peut donc s'exprimer à l'aide du R^2 qu'on appelle *coefficient de détermination de la régression*

Le test d'hypothèse linéaire $\mathrm{H}_0 : \beta_p = \ldots = \beta_1 = 0$ contre $\mathrm{H}_1 : \exists i > 0,\ \beta_i \neq 0$ a pour région critique, au niveau α,

$$W = \left\{ y \in \mathbb{R}^n \,;\, \frac{n-p-1}{p} \frac{r^2}{1-r^2} > f_{p,n-p-1,\alpha} \right\}$$

où r^2 la valeur critique $f_{p,n-p-1,\alpha}$ est déterminée par $P\left(F > f_{p,n-p-1,\alpha}\right) = \alpha$ et $F \rightsquigarrow \mathcal{F}(p, n-p-1)$.

1.4.3 Tests d'hypothèses linéaires

Définition 8. *Un test d'hypothèses linéaires est un test de* $\mathrm{H}_0 : \boldsymbol{\beta} \in \xi_0$ *contre* $\mathrm{H}_1 : \boldsymbol{\beta} \notin \xi_0$ *où* ξ_0 *est un sous espace-vectoriel de* $\mathbb{R}^{p+1}$.

Théorème 3. *Le test d'hypothèse linéaire* $\mathrm{H}_0 : \boldsymbol{\beta} \in \xi_0$ *contre* $\mathrm{H}_1 : \boldsymbol{\beta} \notin \xi_0$ *a pour région critique, au niveau* α,

$$W = \left\{ y \in \mathbb{R}^n \,;\, \frac{n-p-1}{p+1-q} \frac{\|\Pi_{E_X \cap E_0^\perp} y\|^2}{\|\Pi_{E_X^\perp} y\|^2} > f_{p+1-q,n-p-1,\alpha} \right\}$$

où
- $E_0 = \{\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \xi_0\}$ *est un sous-espace vectoriel de* $\mathbb{R}^n$.
- $q = dim\, \xi_0 = dim\, E_0$.
- $P\left(F_{p+1-q,n-p-1} > f_{p+1-q,n-p-1,\alpha}\right) = \alpha$ *et* $F_{p+1-q,n-p-1} \rightsquigarrow \mathcal{F}(p+1-q, n-p-1)$.

Ce test est parfois appelé test de Fisher ou F-Test.

Cas particulier : Test de pertinence de la régression.

Sous $\mathrm{H}_0 : \beta_p = \ldots = \beta_1 = 0$, le modèle s'écrit $\boldsymbol{Y} = \beta_{0|\mathrm{H}_0}\, \mathbb{1}_n + \varepsilon_{|\mathrm{H}_0}$. Sous H_0, $\hat{\beta}_{0|\mathrm{H}_0} = \overline{y}_n$, la valeur de $\boldsymbol{Y}$ prédite par le modèle est alors $\widehat{\boldsymbol{Y}}_0 = \widehat{\boldsymbol{B}}_{0|\mathrm{H}_0}\, \mathbb{1}_n = \overline{\mathcal{Y}}$.

Comme $q = 1$, on retrouve le test de pertinence de la régression linéaire.

1.5 Moindres carrés généralisés

1.5.1 Résultats généraux

Dans le cas le plus général, la matrice de covariance V des résidus est une matrice symétrique semi-définie positive quelconque.

En fait, il est facile d'obtenir un estimateur sans biais de $\boldsymbol{\beta}$ en ignorant tout de la matrice de covariance du modèle comme le montre la proposition suivante :

Proposition 14. *Pour toute matrice M telle que ${}^t\boldsymbol{X}\,M\,\boldsymbol{X}$ soit inversible, $({}^t\boldsymbol{X}\,M\,\boldsymbol{X})^{-1}{}^t\boldsymbol{X}\,M\,\boldsymbol{Y}$ est un estimateur sans biais de $\boldsymbol{\beta}$.*

La question qui se pose maintenant est de trouver parmi tous ces estimateurs linéaires sans biais de $\boldsymbol{\beta}$ celui de variance minimale.

Proposition 15. *L'estimateur sans biais de variance minimale parmi tous les estimateurs sans biais linéaires est*

$$\widehat{\boldsymbol{B}} = \left({}^t\boldsymbol{X}\,V^{-1}\,\boldsymbol{X}\right)^{-1}\,{}^t\boldsymbol{X}\,V^{-1}\,\boldsymbol{Y}$$

Autrement dit, le choix optimal pour la matrice M est V^{-1}.

Dans la pratique, la matrice de covariance V est rarement connue et il est donc impossible de calculer l'estimateur des

Pour contourner ce problème, il arrive souvent qu'on suppose que la matrice de covariance des résidus est du type $K = \sigma^2\,B$ où σ^2 est un paramètre positif inconnu et B une matrice semi-définie positive connue. La méthode des M.C.G. devient alors la méthode des *moindres carrés pondérés*. Si B n'est pas diagonale, cela signifie qu'il existe des corrélations non nulles entre les résidus.

1.5.2 Moindres carrés pondérés

Dans ce paragraphe, on suppose que $K_{\boldsymbol{\varepsilon}} = \sigma^2\,B$ où B est une matrice semi-définie positive connue. Le principal résultat est le théorème de Gauss-Markov généralisé suivant :

Théorème 4 (de Gauss-Markov généralisé)**.**

1. *L'estimateur des moindres carrés pondérés $\widehat{\boldsymbol{B}}$ de $\boldsymbol{\beta}$ vérifie l'équation normale*

$$ {}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{X}\,\widehat{\boldsymbol{B}} = {}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{Y}$$

2. *Si ${}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{X}$ est inversible, alors $\widehat{\boldsymbol{B}} = ({}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{X})^{-1}\,{}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{Y}$ est l'estimateur sans biais de $\boldsymbol{\beta}$ de variance minimale parmi les estimateurs sans biais linéaires.*
3. *$K_{\widehat{\boldsymbol{B}}} = \sigma^2({}^t\boldsymbol{X}\,B^{-1}\,\boldsymbol{X})^{-1}$*
4. *$\widehat{\sigma}^2 = \frac{1}{n-p-1}{}^t(\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{B}})\,B^{-1}\,(\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{B}})$ est un estimateur sans biais de σ^2.*

1.5.3 Notion de métrique

Soit M une matrice symétrique définie positive d'ordre n. Alors l'équation,

$$\forall\,(x,\,y)\in\mathbb{R}^n,\ \ <x,\,y>_M={}^t x\,M\,y$$

définit un produit scalaire sur $\mathbb{R}^n$.

Le produit scalaire euclidien

$$<x,\,y>=\sum_{i=1}^{n}x_iy_i={}^t xy$$

correspond à $M=I_n$.

La norme associée est $\|x\|_M^2={}^t x\,M\,x$.

La distance associée

$$d_M(x,y)=\|y-x\|_M=\sqrt{{}^t(y-x)\,M\,(y-x)}$$

est appelée une *métrique* de $\mathbb{R}^n$. Par extension, la matrice M est elle-même appelée métrique.

L'estimateur des moindres carrés pondérés de σ^2 peut être réécrit à l'aide de ces définitions :

$$\widehat{\sigma}^2=\frac{1}{n-p-1}\|\boldsymbol{Y}-\boldsymbol{X}\,\widehat{\boldsymbol{B}}\|_{B^{-1}}^2$$

Par ailleurs, l'estimateur $\widehat{\boldsymbol{B}}$ des moindres carrés pondérés de $\boldsymbol{\beta}$ est par définition :

$$\widehat{\boldsymbol{B}}=\mathrm{Arg}\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\|Z-\boldsymbol{X}'\,\boldsymbol{\beta}\|^2=\mathrm{Arg}\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\|\boldsymbol{Y}-\boldsymbol{X}\,\boldsymbol{\beta}\|_{B^{-1}}^2$$

Cela signifie que faire des moindres carrés pondérés revient à changer de distance dans $\mathbb{R}^n$: on passe de la métrique $M=I_n$ pour les M.C.O. appelée *métrique identité* ou *métrique canonique* à la métrique $M=B^{-1}$.

Muni de la métrique M quelconque, $\mathbb{R}^n$ reste un espace euclidien. Il est donc toujours possible d'y définir des projections orthogonales. Pour éviter des confusions, ces projections seront appelées *projections M-orthogonales*. Ainsi :

— le projecteur M-orthogonal sur E_X est

$$\Pi_{E_X}^M=\boldsymbol{X}\,({}^t\boldsymbol{X}\,M\,\boldsymbol{X})^{-1}\,{}^t\boldsymbol{X}\,M$$

— le projecteur M-orthogonal sur $E_X^{\perp}$ est

$$\Pi_{E_X^{\perp}}^M=I_n-\boldsymbol{X}\,({}^t\boldsymbol{X}\,M\,\boldsymbol{X})^{-1}\,{}^t\boldsymbol{X}\,M$$

— $\boldsymbol{Y}=\boldsymbol{X}\widehat{\boldsymbol{B}}$ est la projection M-orthogonale de $\boldsymbol{Y}$ sur E_X.

1.6 Cas des variables catégorielles

Considérons le cas d'une (seule) variable catégorielle. C'est à dire les colonnes $x_{,1}$, ..., $x_{,j}$ sont issues d'un codage d'une variable catégorielle. Il existe un nombre fini de valeurs (observées pour simplifier) $v_1, \dots, v_p$, aussi appelé modalité. Naturellement ces valeurs non réelles doivent être codées d'une façon ou d'une autre.

1.6.1 Cas de variables non ordonnées

On choisit de coder les modalités

$$x_{i,j} = 1_{x_i = v_j}.$$

Si les modalités sont uniques et exclusives, alors pour toute observation i la ligne $x_{i,\cdot}$ possède exactement un seul coefficient à 1 et tous les autres à zéro. Le recodage est donc

$$\begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \Leftrightarrow \begin{pmatrix} x_{,1} & \dots & x_{,p} \\ 1 & 0 & \\ & & \\ 0 & \ddots & 0 \\ & 0 & 1 \end{pmatrix}.$$

Voici un exemple de 10 données de `chickwts` en tableau 1.1.

	chickwts$feed	feed=casein	feed=horsebean	feed=linseed	feed=meatmeal	feed=soybean	feed=sunflower
55	meatmeal	0	0	0	1	0	0
7	horsebean	0	1	0	0	0	0
60	casein	1	0	0	0	0	0
51	meatmeal	0	0	0	1	0	0
1	horsebean	0	1	0	0	0	0
69	casein	1	0	0	0	0	0
13	linseed	0	0	1	0	0	0
26	soybean	0	0	0	0	1	0
45	sunflower	0	0	0	0	0	1
8	horsebean	0	1	0	0	0	0

TABLE 1.1 – recodage de la variable `feed` – `chickwts`

Comme chaque ligne a une somme égale à 1, si on introduit un coefficient constant unitaire alors la matrice suivante n'est pas de plein rang puisque la première colonne est la somme des autres colonnes.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}.$$

Par conséquent il est nécessaire d'imposer des contraintes d'identifiabilité : $R\boldsymbol{\beta} = 0$ où R est la matrice de contrasts. Voici les 3 exemples usuels

- modalité de référence : $R = (0, 1, 0, \dots, 0)$ implique de supprimer la deuxième colonne et $\beta_1 = 0$. En R, c'est le comportement par défaut. Voir l'exemple en tableau 1.2.
- absence de terme constant : $R = (1, 0, \dots, 0)$ implique de supprimer la première colonne et $\beta_0 = 0$. En R, il suffit de préciser `0+` dans la formule. Voir l'exemple en tableau 1.3.
- condition zéro-somme : $R = (0, 1, \dots, 1)$ entraine que $\sum_{i=1}^{p} \beta_i = 0$. En R, il faut l'argument `contr=contr.sum`. Voir l'exemple en tableau 1.4.

	(Intercept)	feed=horsebean	feed=linseed	feed=meatmeal	feed=soybean	feed=sunflower
55	1	0	0	1	0	0
7	1	1	0	0	0	0
60	1	0	0	0	0	0
51	1	0	0	1	0	0
1	1	1	0	0	0	0
69	1	0	0	0	0	0
13	1	0	1	0	0	0
26	1	0	0	0	1	0
45	1	0	0	0	0	1
8	1	1	0	0	0	0

TABLE 1.2 – Modalité de référence (`casein`)

	feed=casein	feed=horsebean	feed=linseed	feed=meatmeal	feed=soybean	feed=sunflower
55	0	0	0	1	0	0
7	0	1	0	0	0	0
60	1	0	0	0	0	0
51	0	0	0	1	0	0
1	0	1	0	0	0	0
69	1	0	0	0	0	0
13	0	0	1	0	0	0
26	0	0	0	0	1	0
45	0	0	0	0	0	1
8	0	1	0	0	0	0

TABLE 1.3 – Pas de constante

	(Intercept)	feed=casein	feed=horsebean	feed=linseed	feed=meatmeal	feed=soybean	feed=sunflower
55	1	0	0	0	1	0	0
7	1	0	1	0	0	0	0
60	1	1	0	0	0	0	0
51	1	0	0	0	1	0	0
1	1	0	1	0	0	0	0
69	1	1	0	0	0	0	0
13	1	0	0	1	0	0	0
26	1	0	0	0	0	1	0
45	1	0	0	0	0	0	1
8	1	0	1	0	0	0	0

TABLE 1.4 – Toute modalité et constante