Université Paris 7- Denis Diderot

Notes de cours pour

le Module

DATA MINING

Dominique Picard[1]

Chapitre 1

Introduction aux modèles de régression

[1]

Une première citation de H.G. Wells (1866-1946) : 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'

Une deuxième citation de Hal Varian, The McKinsey Quarterly, January 2009 : "I keep saying the sexy job in the next ten years will be statisticians. "

Le modèle de régression est probablement le modèle le plus vaste et le plus utilisé et étudié (encore maintenant des milliers d'articles paraissent dans des revues mathématiques chaque année sur le sujet) en statistique.

Il consiste à proposer une modélisation dans le cas de figure suivant. Pour employer un vocabulaire d'économiste, on dispose d'une variable endogène ou **expliquée** que l'on note généralement Y et d'un certain nombre p de variables exogènes ou **explicatives** que l'on note généralement $X^1, \ldots, X^p$. Les variables X^j pour $j = 1, \ldots, p$ apparaissant comme les causes d'un phénomème et la variable Y comme une conséquence, on a envie d'écrire qu'il existe une relation fonctionnelle entre la variable Y et les variables X^j pour $j = 1, \ldots, p$ soit

$$Y = f(X^1, \ldots, X^p)$$

pour une certaine fonction f sur laquelle on veut avoir des informations.

Le but de ce cours est d'étudier les principales méthodes d'estimation de cette fonction f lorsqu'on dispose de n données sur les variables $Y, X^1, \ldots, X^p$. Nous serons amenés à distinguer le cas où $p < n$ du cas où $p >> n$, plus difficile mais aussi très important dans le cadre actuel marqué plutot par la surabondance des données.

Suivant les hypothèses que l'on est pret à faire a priori, plusieurs méthodes seront envisagées. En particulier nous envisagerons plusieurs types de modèles.

- le modèle linéaire : f est supposée être une fonction linéaire. On s'intéresse à des variables X^j quantitatives.
- L'ANOVA ou le modèle logistique : f est supposée être linéaire. On s'intéresse à des variables X^j qualitatives ou catégorielles.
- le modèle de classification où les variables X^j sont quantitatives et où la variable Y est qualitative.
- le modèle paramétrique : f est supposée dépendre d'un paramètre θ inconnu. Mais la forme $f := f_\theta$ est connue.
- le modèle non paramétrique : f est supposée être complêtement inconnue. Cependant, on suppose qu'elle admet une certaine régularité.

Ces modèles sont très utilisés dans la pratique et dans de nombreux domaines. Donnons quelques exemples.

- Dans le domaine de l'économie : En vue d'une politique de relance par la consommation, on veut connaitre l'influence du revenu sur la consommation. Soit R le revenu d'un ménage et C sa consommation. L'INSEE modélise généralement la relation entre R et C par un modèle linéaire

$$R = a + bC.$$

  Le paramètre a représente la consommation incompressible d'un ménage (même sans revenu) et le paramètre b est appelé la **propension marginale à consommer**. Une estimation de b proposée par l'INSEE est environ 0.8.
- Dans le domaine de la biomédecine : On veut évaluer le risque d'apparition d'un cancer selon que la personne a été (ou non) exposée au tabac. La variable explicative est ici X qui prend 2 valeurs ("Fumeur" ou "NonFumeur") et la variable à expliquer est Y qui est une probabilité de risque (valeur comprise entre 0 et 1). On propose comme modèle

$$\text{Logit}(Y) = a + bX.$$

- Dans le domaine de l'environnement : il s'agit de prévoir la concentration d'ozone à partir des variables suivantes : force du vent, température et concentration d'oxyde d'azote. La forme particulière de la fonction f_θ est donnée par des physiciens qui utilisent des équations provenant de la mécanique des fluides.
- En signal : On enregistre un concert. On discrétise le signal en échantillonnant toutes les secondes. On note Y_i le signal recu au temps i. Ce signal est fonction du temps et on modélise par

$$Y_i = f(i) + \epsilon_i$$

  où ϵ_i contient tous les "bruits" enregistrés mais indésirables (les toux des gens, le bruit de la ventilation, ect..).

Chapitre 2

Modèle de régression linéaire

[1]

Ce modèle de régression est le plus utilisé et le mieux connu de toutes les personnes traitant des données dans des domaines divers.

2.1 Description du modèle

Soit Y la variable que l'on veut expliquer grâce aux p variables explicatives $X^1, \dots X^p$. On note X la matrice $n \times p$ qui contient les échantillons des variables X^j pour $j = 1, \dots p$:

$$X = \left(X_i^j\right)_{1 \le i \le n, 1 \le j \le p}.$$

La modélisation dite de régression linéaire multiple est la suivante

$$Y_i = \beta_1 \, X_i^1 + \ldots + \beta_p \, X_i^p + \epsilon_i, \quad 1 \le i \le n$$

ce qui est équivalent, en écriture matricielle à

$$\begin{array}{ccccccc} Y & = & X & \beta & + & \epsilon \\ (n,1) & & (n,p) & (p,1) & & (n,1) \end{array} \tag{2.1}$$

avec :

1. β est un paramètre de $\mathcal{R}^p$ inconnu et non aléatoire.
2. on impose au vecteur aléatoire ϵ de $\mathcal{R}^N$:
   - centrage : $E(\epsilon) = 0_n$.
   - indépendance et homoscédasticité : notons Σ la matrice de variance-covariance de ϵ. Alors $\Sigma = \sigma^2 Id_n$ pour $\sigma^2 > 0$ inconnu, déterministe.

   On ne connait pas forcément la loi de ϵ. On appelle ϵ l'erreur ou la perturbation.

Remarquons qu'en géneral, la constante 1_n de R^n fait partie des régresseurs (par défaut dans les logiciels). Le modèle est dit linéaire car il est linéaire en les paramètres β_j pour $j = 1, \dots, p$.

La plupart du temps dans ce cours, nous ferons l'hypothèse que les ε_i sont i.i.d. de loi normale $N(0, \sigma^2)$.

Une fois la modélisation choisie, il s'agit d'estimer les paramètres inconnus β, σ^2 du modèle (il y en a donc $p + 1$ au total).

2.1.1 Exemples

1. Comparaison de 2 populations de même variance : On dispose de 2 échantillons $Z_1, \ldots, Z_m$ i.i.d. $N(\mu_1, \sigma^2)$ et $Z'_1, \ldots, Z'_m$ i.i.d. $N(\mu_2, \sigma^2)$. On les concatène pour former le vecteur

$$Y = (Z_1, \ldots, Z_n, Z'_1, \ldots, X'_m)^* = (Y_1, \ldots, Y_{m+n})^*$$

Si on considère la matrice X de taille $n \times 2$, telle que

$$X_1^1 = \ldots = X_n^1 = 1,\ X_{n+1}^1 = \ldots = X_{n+m}^1 = 0$$

$$X_1^2 = \ldots = X_n^2 = 0,\ X_{n+1}^2 = \ldots = X_{n+m}^2 = 1$$

et le vecteur $\beta = (\mu_1, \mu_2)^*$, il est facile de mettre notre modèle sous la forme (2.1).

2. Droite de régression. Supposons que l'on sache par des arguments théoriques (agronomiques, biologiques, économiques, physiques,...) que 2 quantités x (par exemple le temps) et y (par exemple la taille d'un animal) sont liées par une équation affine de la forme $y = ax + b$, dont on veut identifier les coefficients a et b. Une façon de procéder est de mesurer y_i pour différentes valeurs de x_i (appelée variable contrôlée) et de modèliser les erreurs par des $N(0, \sigma^2)$ indépendantes. On a alors la représentation (2.1), avec

$$X_1^1 = x_1, \ldots, X_n^1 = x_n,$$

$$X_1^2 = \ldots = X_n^2 = 1,$$

$$\beta = (a, b)^*$$

Cet exemple peut se généraliser en remplaçant la relation affine par une relation de la forme :

$$y = \sum_{j=0}^{p} \beta_j f_j(x)$$

Une régression polynomiale s'obtient par exemple en prenant

$$f_0 = 1,\ f_1(x) = x, \ldots, f_p(x) = x^p$$

3. On appelle **Analyse de la variance** (Anova) le cas où la matrice X est uniquement constituée de 1 et de 0.

Donnons un exemple : Dans des conditions de culture de référence (0), une variété de blé a un rendement moyen de μ. On la soumet, dans des parcelles expérimentales à un traitement à 2 facteurs :

1er facteur (par exemple, un engrais) auquel, outre le niveau 0 de référence, on donne 2 niveaux, notés 1 et 2 (par exemple, 2 doses différentes d'engrais).

2eme facteur (par exemple, un niveau d'ensoleillement) auquel on donne soit le niveau de référence 0 soit le niveau 1.

Le modèle de base choisi est le suivant :

$$y = \mu + \alpha_i + \beta_j$$

Il est dit additif : Le effets des facteurs s'ajoutent simplement sans interférences. α_i représente l'effet du 1er facteur au niveau $i = 0, 1, 2$, β_j représente l'effet du 2eme facteur au niveau $j = 0, 1$. $\alpha_0 = \beta_0 = 0$. Le terme additif signifie que les effets des 2 facteurs s'ajoutent. Il est clair qu'on aurait pu aussi rajouter "une interaction" de la forme γ_{ij}, mais par souci de simplicité, nous ne l'avons pas fait ici.

Le but est d'obtenir des informations (estimation ou test) sur les α_i et les β_j. Pour cela, on réalise une expérimentation : On divise un champs en parcelles numérotées (6, dans l'exemple qui suit). Sur chaque parcelle, on applique les facteurs à un niveau prescrit. La description des niveux affectés aux parcelles s'appelle le plan de l'expérience. Ici, il est donné par le tableau suivant.

Parcelle	1	2	3	4	5	6
Facteur 1	0	1	2	0	1	0
Facteur 2	0	0	0	0	0	1

Si l'on suppose que l'on modèlise le rendement sur chaque parcelle par un effet de type (3) auquel s'ajoute une erreur $N(0, \sigma^2)$, et si l'on suppose les erreurs indépendantes, on obtient une équation du type $Y = X\beta + \varepsilon$, où Y est le vecteur des rendements, ε est le vecteur des erreurs, $\beta = (\mu, \alpha_1, \alpha_2, \beta_1)^*$ et X est la matrice suivante

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

2.2 Méthode des Moindres Carrés Ordinaires

2.3 Estimation de β

Nous allons utiliser ici la méthode dite des moindres carrés : Pour cela, on introduit la fonction,

$$\gamma(\beta, Y) = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2$$

Cette fonction mesure la distance dans $\mathbb{R}^n$ entre le vecteur Y et sa prédiction par $X\beta$. Il est relativement naturel de choisir comme estimateur de β, un point $\hat{\beta}$ rendant cette quantité minimum.

$$\hat{\beta} = \text{Argmin}\{\gamma(\beta, Y); \beta \in \mathbb{R}^p\}$$

2.3.1 Interprétation géométrique

Si β parcourt $\mathbb{R}^p$, $X\beta$ parcourt l'espace vectoriel V engendré, dans $\mathbb{R}^n$, par les colonnes de la matrice X :

$$V = X(\mathbb{R}^p) \subset \mathbb{R}^n$$

Comme $\gamma(\beta, Y) = \|Y - X\beta\|^2$, nécessairement $X\hat{\beta}$, existe, est unique puisque c'est la projection sur V de Y, $X\hat{\beta} = \mathrm{Proj}_V(Y)$. On en déduit que $\hat{\beta}$ existe aussi toujours, mais n'est unique que si X est injectif.

Proposition 1 *Si $p \leq n$, la matrice X, de dimension $n \times p$ est injective si et seulement si X^*X est inversible.*

Démonstration de la Proposition.

Il sufit de démontrer que $\ker(X) = \ker(X^*X)$. Il est clair que $\ker(X) \subset \ker(X^*X)$. Maintenant, soit $u \in \ker(X^*X)$, on a $X^*Xu = 0$, d'où $u^*X^*Xu = 0$, i.e. $\|Xu\|^2 = 0 \Longrightarrow Xu = 0 \Longrightarrow u \in \ker X$.

Résolution algébrique

$$\begin{array}{rcll} X\hat{\beta} = \mathrm{Proj}_V(Y) & \Longleftrightarrow & \langle Y - X\hat{\beta}, Xb\rangle = 0, & \forall b \in \mathbb{R}^p \\ & \Longleftrightarrow & b^*X^*Y = b^*X^*X\hat{\beta}, & \forall b \in \mathbb{R}^p \\ & \Longleftrightarrow & X^*Y = X^*X\hat{\beta} & \end{array}$$

D'oú, en utilisant la proposition si X est injective,

$$\hat{\beta} = (X^*X)^{-1}X^*Y$$

Remarque : Si X^*X n'est pas inversible, on n'a pas unicité de $\hat{\beta}$, mais existence. Donnons une solution, utilisant la pseudoinverse : X^*X étant une matrice symétrique, positive, elle s'écrit M^*DM avec M matrice orthogonale et D est une matrice diagonale, dont les coefficients diagonaux sont notés r_i^2. On suppose $r_i^2 > 0, \forall i = 1, \ldots, k,\ r_i^2 = 0, \forall i \geq k+1$. Appelons pseudoinverse de X^*X la matrice

$$(X^*X)^{(-1*)} = M^* \begin{pmatrix} \frac{1}{r_1^2} & \ldots & \ldots & 0 & 0 & 0 \\ & & \ldots & & & \\ 0 & \ldots & \frac{1}{r_k^2} & \ldots & 0 & 0 \\ & & \vdots & & & \\ 0 & \ldots & 0 & \ldots & 0 & 0 \end{pmatrix} M$$

Notons que si X^*X est inversible, alors pseudoinverse et inverse coincident. On vérifie facilement que

$$\hat{\beta} = (X^*X)^{(-1*)}X^*Y$$

est une solution de notre problème, et que l'opérateur de projection sur V est donné par :

$$X\hat{\beta} = X(X^*X)^{(-1*)}X^*Y = \mathrm{Proj}_V(Y)$$

$\triangle$

Rappelons que si $V^\perp$ est le supplémentaire orthogonal de V,

$$\mathrm{Proj}_{V^\perp}(Y) = Y - \mathrm{Proj}_V(Y) = [I_n - \mathrm{Proj}_V](Y) = [I_n - X(X^*X)^{-1}X^*]Y$$

Définition 1 *On appelle vecteur des résidus, le vecteur*

$$\hat{\varepsilon} = [I_n - X(X^*X)^{-1}X^*]Y.$$

Il représente l'erreur de prédiction. Le carré de sa norme s'appelle l'erreur quadratique.

Exemples :

1. Dans le cas élémentaire suivant :

$$Y_i = \mu + \varepsilon_i$$

l'estimateur des moindres carrés se calcule facilement et vaut $\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$.

2. Dans le cas d'une régression linéaire, nous avons vu que $\beta = (a, b)^*$ et

$$X = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

De sorte que

$$X^*X = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix}$$

Dans ce cas, un changement de paramètres peut rendre les choses plus aisées : En effet, si on introduit $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$, le modèle s'ecrit :

$$Y_i = az_i + b' + \varepsilon_i,\ z_i = x_i - \bar{x}_n,\ b' = b + \bar{x}_n$$

et clairement minimiser $\sum_{i=1}^n (Y_i - az_i + b')^2$ équivaut à minimiser $\sum_{i=1}^n (Y_i - ax_i + b)^2$, avec la relation suivante $\hat{b}' = \hat{b} + \hat{a}\bar{x}_n$. L'équation (2) introduit un nouveau modèle linéaire dont la matrice X' s'écrit :

$$X'^*X' = \begin{pmatrix} \sum_{i=1}^n z_i^2 & 0 \\ 0 & n \end{pmatrix}$$

Cette matrice est inversible si et seulement si $\sum_{i=1}^n z_i^2 \neq 0$, c'est à dire si les x_i ne sont pas tous égaux. Dans ce cas, on obtient facilement :

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)Y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2},\ \hat{b} = \bar{Y}_n + \hat{a}\bar{x}_n$$

3. Considérons maintenant la régression périodique suivante :

$$Y_i = a_0 + a_1 \cos(2\pi \frac{i}{n}) + a_2 \sin(2\pi \frac{i}{n}) + \varepsilon_i,\ i = 1, \ldots, n$$

On vérifie que en utilisant les relations sur les racines de l'unité que X^*X se met sous la forme suivante :

$$\begin{pmatrix} n & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n})^2 & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n})^2 \end{pmatrix} = \begin{pmatrix} n & 0 & 0 \\ 0 & \frac{n}{2} & 0 \\ 0 & 0 & \frac{n}{2} \end{pmatrix}$$

On en déduit que

$$\hat{a}_0 = \bar{Y}_n,\ \hat{a}_1 = \sum_{i=1}^n \cos(2\pi \frac{i}{n}) Y_i,\ \hat{a}_2 = \sum_{i=1}^n \sin(2\pi \frac{i}{n}) Y_i$$

△

2.3.2 Calcul récursif, Méthode de Gram Schmidt

Nous proposons ici une méthode pour calculer $\hat{\beta}$ de façon récursive. Appelons X^j la colonne numéro j de la matrice X pour $1 \leq j \leq p$.

Considérons le cas suivant dans lequel les MCO sont particulièrement faciles à calculer : Supposons que les **colonnes** de X soient orthogonales (i.e. X^tX est une matrice diagonale dont les coeficients diagonaux sont les carrés des normes des colonnes : $\sum_{i=1}^n [X_i^j]^2 = \langle X^j, X^j \rangle$. Dans ce cas, les coefficients $\hat{\beta}_j$ valent simplement :

$$\hat{\beta}_j = \frac{\langle X^j, Y \rangle}{\langle X^j, X^j \rangle}$$

Rappelons nous maintenant le procédé d'orthonormalisation de Gram Schmidt qui pour des vecteurs quelconques $u_1, \ldots, u_k$ (tels que l'espace engendré par ces vecteurs ($sp\ \{u_1, \ldots, u_k\}$) soit de dimension k) introduit les vecteurs $v_1, \ldots, v_k$ qui sont orthogonaux et vérifient $sp\ \{u_1, \ldots, u_l\} = sp\ \{v_1, \ldots, v_l\}$, pour tout $1 \leq l \leq k$. Ce procédé consiste simplement à construire les v_l sous la forme suivante : $v_1 = u_1$,

$$v_\ell = u_\ell - P_{v_{\ell-1}} u_\ell - \ldots - P_{v_1} u_\ell, \quad \ell \geq 2.$$

(P_{v_j} désigne la projection sur le vecteur v_j).

Remarquons que pour $1 \leq j \leq \ell - 1$,

$$P_{v_j} u_\ell = \frac{\langle v_j, u_\ell \rangle}{\langle v_j, v_j \rangle}.$$

De plus comme les v_j sont orthogonaux, $P_{v_{\ell-1}} u_\ell + \ldots + P_{v_1} u_\ell$ est la projection de u_ℓ sur l'espace $sp\{v_1, \ldots, v_{\ell-1}\}$. Donc v_l est en fait le 'résidu' de la projection de la projection de u_ℓ sur l'espace $sp\{v_1, \ldots, v_{\ell-1}\}$.

Considérons maintenant, dans le cas $p \leq n$ et où la matrice X est de rang p, l'algorithme suivant :

– Initialisation : $Z^1 = X^1$
– Pour $l = 2$ jusqu'à p calculer : Z^l le résidu de la projection de X^l sur $Z^{l-1}, \ldots, Z^1$, i.e.

$$Z^l = X^l - \frac{\langle Z^{l-1}, X^l \rangle}{\langle Z^{l-1}, Z^{l-1} \rangle} Z^{l-1} - \ldots - \frac{\langle Z^1, X^l \rangle}{\langle Z^1, Z^1 \rangle} Z^1.$$

Montrer qu'alors

$$\hat{\beta}_p = \frac{\langle Z^p, Y \rangle}{\langle Z^p, Z^p \rangle}.$$

En changeant l'ordre des colonnes de la matrice X, on peut s'arranger pour faire apparaitre X^j en dernier pour chaque j. Cela donne une façon de calculer les $\hat{\beta}_j$ sans inverser la matrice. (Attention on a donc p calculs différents.)

Cet algorithme permet aussi de mesurer les problèmes qui peuvent arriver au cours d'une telle estimation. Supposons en effet que le vecteur X^p soit très corrélé avec (par exemple) X^{p-1} (ou soit proche d'une combinaison linéaire de $X^1, \ldots, X^{p-1}$) ; dans ce cas le résidu Z_p va être très petit et par voie de conséquence l'estimation de $\hat{\beta}_p$ très instable.

2.4 Lois des estimateurs. Estimation de σ^2.

Nous allons maintenant montrer la proposition suivante sous l'hypothèse que les ε_i sont i.i.d. $N(0, \sigma^2)$:

Proposition 2 *Sous la condition, $p \leq n$, X^*X inversible, le vecteur de dimension $p + n$:*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix}$$

est un vecteur gaussien de moyenne et variance :

$$\begin{pmatrix} \beta \\ 0 \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} (X^*X)^{-1} & 0 \\ 0 & I_n - X(X^*X)^{-1}X^* \end{pmatrix}$$

Preuve de la Proposition

Espérances et variances de $\hat{\beta}$ Dans ce paragraphe, l'hypothèse de gaussiannité sur les ε_i est inutile. Les résultats sont encore vrais si l'on suppose que $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}\varepsilon = \sigma^2 I_n$.

Comme $\hat{\beta} = (X^*X)^{-1}X^*Y$, on a $\mathbb{E}\hat{\beta} = \mathbb{E}(X^*X)^{-1}X^*(X\beta + \varepsilon) = \beta$.

D'autre part,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X^*X)^{-1}X^*[\mathrm{Var}(Y)]X(X^*X)^{-1} \\ &= (X^*X)^{-1}X^*[\mathrm{Var}(\varepsilon X)](X^*X)^{-1} \\ &= \sigma^2(X^*X)^{-1}X^*X(X^*X)^{-1} = \sigma^2(X^*X)^{-1}. \end{aligned}$$

Loi du vecteur Le vecteur

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix}$$

est fonction linéaire du vecteur Y, c'est donc un vecteur gaussien. Nous avons calculé la moyenne de $\hat{\beta}$ au paragraphe précédent. Il est immédiat que $\mathbb{E}\hat{\varepsilon} = 0$. Nous avons vu que : $X\hat{\beta} = \mathrm{Proj}_V(Y) = X\beta + e$ avec $e = \mathrm{Proj}_V(\varepsilon)$.

De plus, $\hat{\varepsilon} = [I_n - \mathrm{Proj}_V](Y) = \mathrm{Proj}_{V^\perp}(Y) = \mathrm{Proj}_{V^\perp}(\varepsilon) = \varepsilon - e$.

Soit maintenant $P_1 = \mathrm{Proj}_V = X(X^*X)^{-1}X^*$ et $P_2 = \mathrm{Proj}_{V^\perp} = I_n - X(X^*X)^{-1}X^*$. On a donc $X\hat{\beta} = X\beta + P_1\varepsilon,\ \hat{\varepsilon} = P_2\varepsilon$.

Par ailleurs, $P_1 + P_2 = I_n$, $\mathrm{rg}(P_1) = \dim V = \mathrm{rg}X = p,\ \mathrm{rg}(P_2) = n - p$. On peut donc appliquer le th'eorème de Cochran et en déduire que e et $\hat{\varepsilon}$ sont indépendants. Par conséquent, $X\hat{\beta}$ et $\hat{\varepsilon}$ sont indépendants. Il en est de même pour $X^*X\hat{\beta}$ et $\hat{\varepsilon}$, et donc pour $\hat{\beta}$ et $\hat{\varepsilon}$. Il nous reste à calculer la matrice de covariance du vecteur $\hat{\varepsilon}$. Mais, comme $\hat{\varepsilon} = P_2\varepsilon$, elle est égale à $\sigma^2 P_2$. Ceci achève la preuve de la proposition. ∎

Estimation de σ^2. En appliquant le résultat de la Proposition 3, nous avons : $\|\hat{\varepsilon}\|^2$ suit une loi $\sigma^2\chi^2(n-p)$. En conséquence, $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p}$ est d'espérance σ^2. C'est donc un estimateur assez naturel de σ^2.

Construction de nouvelles 'erreurs' A partir des résidus on peut construire des nouvelles variables $\eta_1, \ldots, \eta_{n-p}$ qui, elles sont i.i.d.$N(0, \sigma^2)$(et indépendantes de $\hat{\beta}$) :

La matrice P_2 est une matrice de projection orthogonale, donc $P_2 = P_2^* = P_2^2$, de plus c'est une matrice positive. Donc il existe une matrice orthogonale U ($UU^* = U^*U = I_n$), telle que

$$P_2 = U^*DU$$

où D est une matrice diagonale telle que ses coefficients diagonaux valent 1 jusqu'à $rang(P_2) = n - p$ et 0 ensuite. Il est facile de voir que le vecteur $Z = U\hat{\varepsilon} = DUY = DU\varepsilon$ suit une loi $N(0, \sigma^2 D)$, ce qui signifie que $Z_{n-p+1} = \ldots = Z_n = 0$ et si l'on pose $\eta_1 = Z_1, \ldots, \eta_{n-p} = Z_{n-p}$ les η_i sont les nouvelles erreurs cherchées : i.i.d. $N(0, \sigma^2)$ (et indépendantes de $\hat{\beta}$).

En résumé :

$\hat{\beta} \sim N(\beta, \sigma^2(X^*X)^{-1}),\ \hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$
De plus ces 2 estimateurs sont indépendants.

2.5 Théorème de Gauss Markov et Moindres Carrés pondérés.

Considérons le modèle suivant :

$$Y = X\beta + \mathcal{E}$$

où $\mathcal{E}$ est un vecteur gaussien centré, de matrice de covariance $\sigma^2 G$. G est une matrice symétrique définie positive, connue. Un exemple est la matrice

$$G = \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ & \vdots & & \\ 0 & 0 & \dots & v_n \end{pmatrix},$$

qui correspond au fait que les observations sont encore indépendantes mais chaque observation est entachée d'une variance propre (cas hétéroscédastique).

La question que l'on se pose est doit-on, dans ce cas conserver l'estimateur de β, $\hat{\beta} = (X^*X)^{-1}X^*Y$?

La question se pose avec d'autant plus d'acuité qu'un autre estimateur peut sembler tout aussi naturel : En effet, on peut assez simplement transformer le modèle (2.5) en modèle linéaire ordinaire $Z = X'\beta + \varepsilon$: En posant $G = BB^*,\ Z = B^{-1}Y, X' = B^{-1}X,\ \varepsilon = B^{-1}\mathcal{E}$. Dans ce nouveau modèle, on peut calculer l'estimateur usuel des moindres carrés (on remarque en particulier que du fait que G est définie symétrique positive, B est inversible) :

$$\tilde{\beta} = (X'^*X')^{-1}X'^*Z = (X^*G^{-1}X)^{-1}X^*B^{-1*}B^{-1}Y = (X^*G^{-1}X)^{-1}X^*G^{-1}Y.$$

Remarques :

1. Remarquons que par définition, cet estimateur rend minimale la quantité :

$$\|B^{-1}Y - B^{-1}X\beta\|^2 = (Y - X\beta)^*G^{-1}(Y - X\beta)$$

qui représente la norme du vecteur $Y - X\beta$, dans la norme G^{-1}, d'où le nom donné à cet estimateur de moindres carrés pondérés.

Si on considère le cas particulier où G est diagonale, on doit minimiser l'expression

$$\sum_{i=1}^{n} \frac{1}{v_i^2}(Y_i - (X\beta)_i)^2$$

qui tient compte de la crédibilité de chaque observation en raison inverse de sa variance.

2. $\mathrm{Var}(a^*\tilde{\beta}a) = a^*(X^*G^{-1}X)^{-1}a$..
3. Une autre façon d'énoncer la remarque 1 est d'observer que

$$P_V^G = X(X^*G^{-1}X)^{-1}X^*G^{-1}$$

est la matrice associée à l'opérateur de projection dans V, défini avec la métrique G^{-1}. (Rappelons que si A est une matrice symétrique définie positive de $\mathbb{R}^n$, x^*Ay définit un produit scalaire sur $\mathbb{R}^n$ et on peut donc considérer la métrique associée.)

Remarquons que dans ce cas les relations matricielles $P_V = P_V^*,\ P_V^2 = P_V,\ I_n = P_V + P_{V^\perp}$ valides en métrique euclidienne doivent être remplacées par

$$P_V^G = G(P_V^G)^* G^{-1},\ (P_V^G)^2 = P_V^G,\ I_n = P_V^G + P_{V^{\perp,G}}^G. \tag{2.2}$$

où $V^{\perp,G}$ désigne le supplémentaire orthogonal de V, pour le produit scalaire G^{-1}. Ces relations se démontrent à partir des relations classiques en observant que

$$\|x\|_{G^{-1}}^2 = x^* B^{-1*} B^{-1} x = \|B^{-1}x\|_{I_n}^2.$$

On en déduit facilement que

$$P_V^G = B P_{B^{-1}V} B^{-1},\ V^{\perp,G} = B(B^{-1}V)^\perp$$
$$P_{V^{\perp,G}}^G = B P_{(B^{-1}V)^\perp} B^{-1}$$

△

Nous allons montrer que cet estimateur possède en fait des propriétés d'optimalité très intéressantes :

Définition 2 *L'estimateur $\bar{\beta}$ est dit linéaire s'il existe une matrice A telle que $\bar{\beta} = AY$.*

Théorème 1 *Considérons le modèle $Y = X\beta + \mathcal{E}$ où $\mathcal{E}$ est un vecteur aléatoire centré, de matrice de covariance $\sigma^2 G$. G est une matrice symétrique définie positive, connue. Si $\bar{\beta}$ est un estimateur linéaire, tel que $\mathbb{E}_\beta \bar{\beta} - \beta = 0,\ \forall \beta \in \mathbb{R}^p$, Alors, il existe R matrice symétrique positive de $\mathbb{R}^p$, telle que $Var(\bar{\beta}) = Var(\tilde{\beta}) + R$.*

Remarque : La signification de ce théorème, est que $\forall a \in \mathbb{R}^p$, $\mathrm{Var}(a^*\bar{\beta}a) \geq \mathrm{Var}(a^*\tilde{\beta}a)$. Or cette inégalité est très importante, en particulier si le vecteur $\mathcal{E}$ est gaussien et que l'on veut construire un intervalle de confiance. En suivant la démarche du paragraphe suivant, on montre très facilement que dans le cas σ connu, cet intervalle est

$$[a^*\bar{\beta} - z_{\alpha/2}\sqrt{\mathrm{Var}(a^*\bar{\beta}a)}\sigma, a^*\bar{\beta} + z_{\alpha/2}\sqrt{\mathrm{Var}(a^*\bar{\beta}a)}\sigma]$$

si on utilise $\bar{\beta}$ et

$$[a^*\tilde{\beta} - z_{\alpha/2}\sqrt{\mathrm{Var}(a^*\tilde{\beta}a)}\sigma, a^*\tilde{\beta} + z_{\alpha/2}\sqrt{\mathrm{Var}(a^*\tilde{\beta}a)}\sigma]$$

si on utilise $\tilde{\beta}$. Il est clair qu'on a intérêt à prendre la seconde solution puisque la longueur de l'intervalle est plus petite. △

Preuve :

Remarquons d'abord que la condition $\mathbb{E}_\beta \bar{\beta} - \beta = 0,\ \forall \beta \in \mathbb{R}^p$, se traduit encore par $(AX - I_n)\beta = 0,\ \forall \beta \in \mathbb{R}^p$, c'est à dire $AX = I_n$.

Par ailleurs, $\mathrm{Var}(\bar{\beta}) = AGA^*$. Mais on a $I_n = P_V^G + P_{V^{\perp,G}}^G$, en utilisant (2.2). On en déduit :

$$\begin{aligned}\mathrm{Var}(\bar{\beta}) &= A(P_V^G + P_{V^{\perp,G}}^G)GA^* \\ &= AX(X^*G^{-1}X)^{-1}X^*G^{-1}GA^* + AP_{V^{\perp,G}}^G GA^* \\ &= AX(X^*G^{-1}X)^{-1}X^*A^* + R \\ &= \mathrm{Var}(\tilde{\beta}) + R\end{aligned}$$

On finit la démonstration en remarquant que

$$R = AP_{V^{\perp,G}}^G GA^* = ABP_{B^{-1}V^\perp}B^{-1}BB^*A^* = ABP_{B^{-1}V^\perp}B^*A^*$$

Cette quantité est bien symétrique et positive par les propriétés de la projection en métrique euclidienne.

△

2.6 Etude du modèle ajusté : estimation et tests

2.6.1 Intervalles de confiance pour $a^*\beta$ et σ^2

Soit a^* un vecteur de $L(\mathbb{R}^p, \mathbb{R})$, on se propose d'estimer $a^*\beta$.
Exemples :

1. Si $a^* = (1, 0, \ldots, 0)$, on s'intéresse à estimer β_1.
2. Dans l'exemple d'une comparaison de 2 populations, $p = 2$, prendre $a^* = (1, -1)$ consiste à estimer la différence des moyennes. △

On va prendre naturellement $a^*\hat{\beta}$ comme estimateur de $a^*\beta$. Nous nous proposons de construire un intervalle de confiance associé à cette estimation.
Rappel : Supposons que l'on cherche à estimer une quantité $q(\theta)$ réelle.

Définition 3 *Soit α fixé dans $(0,1)$. Soit, dans une expérience arbitraire $\mathcal{E} = (Y, P_\theta, \theta \in \Theta)$, $S = hoY$, $T = h'oY$, 2 estimateurs de $q(\theta)$, on dira que $[S,T]$ est un* **intervalle de confiance au niveau** *α, si*

$$\forall \theta \in \Theta, \quad P_\theta\{q(\theta) \in [S,T]\} \geq 1 - \alpha.$$

Remarque : Bien entendu, $S = -\infty$, $T = \infty$ convient toujours mais n'est guère intéressant. En effet, l'intérêt pratique sera toujours de rendre $T - S$ le plus petit possible.
△

Estimation de $a^*\beta$, σ^2 étant connu

On vérifie que $a^*(\hat{\beta} - \beta) \sim N(0, \sigma^2 a^*(X^*X)^{-1}a)$, de sorte que si $\Phi(z_{\alpha/2}) = \alpha/2$, où

$$\Phi(u) = Prob(\xi \geq u),\ \xi \sim\ N(0,1).$$

$$[a^*\hat{\beta} - z_{\alpha/2}\sqrt{a^*(X^*X)^{-1}a}\sigma, a^*\hat{\beta} + z_{\alpha/2}\sqrt{a^*(X^*X)^{-1}a}\sigma]$$

est un intervalle de confiance pour la quantité $a^*\beta$, au niveau d'erreur α.

Estimation de $a^*\beta$, σ^2 étant inconnu

On a, outre le fait que $a^*(\hat{\beta}-\beta) \sim N(0, \sigma^2 a^*(X^*X)^{-1}a)$, $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$ De plus ces 2 variables aléatoires sont indépendantes. Donc $\frac{a^*(\hat{\beta}-\beta)}{\hat{\sigma}\sqrt{a^*(X^*X)^{-1}a}} \sim T(n-p)$ de sorte que si $\Phi_{n-p}(z_{\alpha/2,n-p}) = \alpha/2$, où

$$\Phi_{n-p}(u) = Prob(\xi \geq u),\ \xi \sim\ T(n-p).$$

$$[a^*\hat{\beta} - z_{\alpha/2}(n-p)\sqrt{a^*(X^*X)^{-1}a}\hat{\sigma}, a^*\hat{\beta} + z_{\alpha/2}(n-p)\sqrt{a^*(X^*X)^{-1}a}\hat{\sigma}]$$

est un intervalle de confiance pour la quantité $a^*\beta$, au niveau d'erreur α.

2.6.2 σ^2

En utilisant le fait que $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$, et la définition de $P(\chi^2(k) > c_{\alpha,k}) = \alpha$, on vérifie facilement que

$$[\frac{\hat{\sigma}^2(n-p)}{c_{\alpha,n-p}}, \frac{\hat{\sigma}^2(n-p)}{c_{1-\alpha/2,n-p}}]$$

est un intervalle de confiance pour la variance au niveau d'erreur α.

2.6.3 Test d'une sous hypothèse linéaire.

Rappel : On se donne un modèle $\mathcal{E} = (Y, P_\theta, \theta \in \Theta)$. On se donne une partition de Θ en deux ensembles (non vides) Θ_0 et Θ_1. Le but du jeu est alors de décider si θ appartient à Θ_0 ou Θ_1.

Définition 4 *Dans le contexte ci-dessus une variable aléatoire $\phi(X)$ à valeurs dans $\{0,1\}$ est appelée* **test**. *La procédure de décision associée consiste à décider Θ_0 si $\phi(x) = 0$ et Θ_1 sinon.*

Notation :
On note généralement :

$$\begin{array}{lll} \mathcal{H}_0, & \text{l'hypothèse 'nulle' :} & \{\theta \in \Theta_0\} \\ \mathcal{H}_1, & \text{'l'alternative' :} & \{\theta \in \Theta_1\} \end{array}$$

Quand on fait un test, il y a deux façon de se tromper, déclarer $\mathcal{H}_1$ alors que $\mathcal{H}_0$ est vrai ou l'inverse. Ceci conduit aux deux définitions suivantes :

Définition 5 *Etant donnée l'epérience $\mathcal{E}$ et le problème de test associé à la partition Θ_0, Θ_1, $\alpha \in [0,1]$, on dit que le test $\phi(X)$ est de* **niveau** *α ssi*

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \phi(X) \leq \alpha$$

Définition 6 *Etant donnée l'expérience $\mathcal{E}$ et le problème de test associé à la partition Θ_0, Θ_1, $\alpha \in [0,1]$, on appelle erreur de deuxième espèce (resp. puissance) la fonction*

$$\theta \in \Theta_1 \mapsto \mathbb{E}_\theta(1-\phi(X)) \quad (resp.\ \mathbb{E}_\theta\phi(X))$$

Nous nous plaçons, comme dans les paragraphes précédents dans le cadre d'un modèle linéaire gaussien, dont la matrice exogène est de rang $p \leq n$. On se donne C, une matrice fixée de dimension $l \times p$, avec $l < p$, on suppose que le rang de C est l et on se propose de tester l'hypothèse $C\beta = 0$.

Exemples :

1. Si $l = 1$, on se ramène à tester la nullit'e d'une forme linéaire. On retrouve donc l'étude du paragraphe précédent.
2. Si par exemple Y_i est la mesure d'un taux de pollution, que l'on cherche à expliquer par différentes variables : X^1 quantité de précipitations, X^2 vitesse du vent, X^3 température, X^4 nombre d'usines, à travers le modèle suivant :
$$Y_i = \beta_1 X_i^1 + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i$$
or, plus modèle contient de paramètres, en général, moins il est interprétable. Donc on peut se poser la question de diminuer le nombre de paramètres, par exemple, en testant $\beta_1 = \beta_3 = 0$. $\triangle$

2.6.4 Résolution

Soit V_1 le sous espace vectoriel de V,

$$V_1 = \{X\beta,\ C\beta = 0\}$$

Comme $\mathrm{rg}(C) = l, \dim(V_1) = \dim(\ker(C)) = p - l$. Soit W_1 le supplémentaire orthogonal de V_1 dans V. On a

$$I_n = P_{V_1} + P_{W_1} + P_{V_\perp},$$

$P_{V_1},\ P_{W_1},\ P_{V_\perp}$ sont des projecteurs respectivement de rang $p-l,\ l,\ n-p$ et donc en appliquant le théorème de Cochran, on a que $(\sigma)^{-1}P_{V_1}\varepsilon,\ (\sigma)^{-1}P_{W_1}\varepsilon,\ (\sigma)^{-1}P_{V_\perp}\varepsilon$ sont des vecteurs gaussiens, indépendants de lois respectives $N(0, P_{V_1}),\ N(0, P_{W_1}),\ N(0, P_{V_\perp})$. D'où, $(\sigma)^{-1}P_{V_1}Y,\ (\sigma)^{-1}P_{W_1}Y,\ (\sigma)^{-1}P_{V_\perp}Y$ sont des vecteurs gaussiens indépendants de lois respectives $N(P_{V_1}X\beta, P_{V_1}),\ N(P_{W_1}X\beta, P_{W_1}),\ N(0, P_{V_\perp})$. On en déduit que :

1. $\|(\sigma)^{-1}P_{V_\perp}Y\|^2 \sim \chi^2(n-p)$.
2. $\|(\sigma)^{-1}P_{V_\perp}Y\|^2$ et $\|(\sigma)^{-1}P_{W_1}Y\|^2$ sont indépendants.
3. – Si $C\beta = 0$, $P_{W_1}(X\beta) = 0$ et donc $\|(\sigma)^{-1}P_{W_1}Y\|^2 \sim \chi^2(l)$.
   – Si $C\beta \neq 0$, $\|(\sigma)^{-1}P_{W_1}Y\|^2 \sim \chi'^2(l, \|P_{W_1}(X\beta))\|^2)$.

On en déduit que sous l'hypothèse $C\beta = 0$, la statistique

$$F = \frac{\|P_{W_1}Y\|^2/l}{\|P_{V_\perp}Y\|^2/(n-p)} \sim F(l, n-p).$$

D'où, si $f_\alpha(n_1, n_2)$, est déterminé par $P(F(n_1, n_2) > f_\alpha(n_1, n_2)) = \alpha$, on a

$$1 - \alpha = P\left(F \in [0, f_\alpha(l, n-p)]\right).$$

Donc,

– Si la statistique F, évaluée sur nos données, tombe en dehors de l'intervalle $[0, f_\alpha(l, n-p)]$, on rejettera l'hypothèse $C\beta = 0$.
– En revanche, si elle tombe dans cet intervalle, on acceptera l'hypothèse.

Ce que l'on vient de décrire s'énonce par la phrase suivante : Le test ϕ qui vaut 1 si $F \geq f_\alpha(l, n-p)$, 0 sinon est un test de niveau α.

2.6.5 Calcul pratique de F

On a

$$F = \frac{\|X\hat{\beta} - P_{V_1}Y\|^2/l}{\|Y - X\hat{\beta}\|^2/(n-p)}$$

–

$$\text{Si } C = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ & & \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix},$$

dans ce cas, on cherche à tester $\beta_1 = \ldots = \beta_l = 0$. Soit $\tilde{X} = (X_{l+1}, \ldots, X_p)$, la matrice des $l-p$ vecteurs colonnes de X. Il est facile de montrer que $P_{V_1}Y = \tilde{X}(\tilde{X}^*\tilde{X})^{-1}\tilde{X}^*Y$, et T se calcule aisément en fonction de X et $\tilde{X}$.

– Dans le cas général, où C est une matrice quelconque, on commence par compléter C en une matrice C' $p \times p$ et inversible, puis on pose $\eta = C'\beta$. Le modèle linéaire $Y = X\beta + \varepsilon$ est équivalent au modèle linéaire suivant, dans lequel on a fait le changement de paramètre $\mu = C'\beta$, $X' = XC'^{-1}$:

$$Y = X'\mu + \varepsilon.$$

Dans ce nouveau modèle l'hypothèse à tester est $\mu_1 = \ldots = \mu_l = 0$ et on est ramené au cas précédent.

2.6.6 Version 'RSS' de ce test

Une autre façon, plus habituelle dans les logiciels d'écrire la statistique F, consiste à introduire les 'sommes des carrés des résidus' dans chaque hypothèse (H_0 et H_1) residuals sum of squares : RSS.
Commençons par H_1, une fois la donnée Y 'expliquée par X, ce qui 'reste à expliquer', les résidus, contribuent pour :

$$RSS1 =: \|Y - X\hat{\beta}\|^2 \ (= \|\hat{\varepsilon}\|^2)$$

De même, sous H_0, la donnée Y est expliquée par $P_{V_1}Y$, donc ce qui 'reste à expliquer' (de façon résiduelle sous H_0) contribue pour :

$$RSS0 =: \|Y - P_{V_1}Y\|^2 = \|Y - X\hat{\tilde{\beta}}\|^2 \ (= \|\hat{\tilde{\varepsilon}}\|^2).$$

Il est clair que $RSS1 \leq RSS0$ et plus précisément, le théorème de Pythagore nous donne :

$$RSS0 - RSS1 = \|P_{W_1}Y\|^2$$

De sorte que l'on peut écrire F sous la forme suivante en introduisant $p_0 =$ dimension sous $H_0 (= p-l$ dans ce qui précède), $p_1 =$ dimension sous $H_1 (=$ p dans ce qui précède) :

$$F = \frac{[RSS0 - RSS1]/(p_1 - p_0)}{RSS1/(n - p_1)}.$$

2.7 Exemples :Etude du modèle ajusté en pratique

Nous allons donner ici des exemples d'utilisation en pratique (et donnés dans les logiciels) des résultats trouvés précédemment.

2.7.1 Significativité globale : le test dit du R^2

Le R^2 en particulier est une quantité à peu près systématiquement donnée dans les logiciels.

Considérons le cas où la constante $1_n = X^1$ fait partie des régresseurs. Pour tester **la significativité globale** du modèle de régression proposé, on peut tester l'hypothèse

$$H_0 : \quad \beta_2 = \beta_3 = \ldots = \beta_p = 0 \qquad \text{contre} \qquad H_1 : \quad \exists j = 2, \ldots p,\ \beta_j \neq 0.$$

Ce qui est bien un test du modèle puisqu'on se demande si on ne ferait pas aussi bien si on ajustait les données simplement par une constante.

Il est clair que $l = p - 1$, $V_1 = sp\{1_n\}$, $P_{V_1} Y = \bar{Y} 1_n$, si $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. Donc

$$RSS0 = \|Y - \bar{Y} 1_n\|^2.$$

Par ailleurs, si on note $\hat{Y} = \sum_{j=1}^p \hat{\beta}_j X^j$, et on a par le théorème de Pythagore,

$$RSS1 = \|Y - \hat{Y}\|^2, \quad RSS0 - RSS1 = \|\hat{Y} - \bar{Y} 1_n\|^2. \tag{2.3}$$

On a donc que la statistique de test s'écrit :

$$F = \frac{n-p}{p-1} \quad \frac{\|\hat{Y} - \bar{Y} 1_n\|^2}{\|Y - \hat{Y}\|^2}.$$

Pour effectuer un test au niveau α, on cherche donc le quantile $q_\alpha = f_\alpha(p - 1, n-p)$ de la loi de Fisher avec les degrés de liberté $p-1, n-p$ et on applique la règle de décision

- si $F > q_\alpha$, H_0 est rejtée et les coefficients ne sont pas globalement nuls. La régression est donc globalement significative.
- si $F \leq q_\alpha$, H_0 est acceptée et les coefficients sont tous nuls. La régression n'est donc pas globalement significative.

Remarque importante $p-$ **value :** Pour résoudre (!?) le problème du choix du niveau du test à prendre ($\alpha = 0.01,\ 0.05,\ 0.1,\ 0.001$ ...?) généralement, les logiciels donnent les $p-$values au lieu des quantiles. La $p-$**value** est par

définiftion le plus petit niveau auquel les données rejetteraient l'hypothèse H_0. En effet si on considère la famille de tests que l'on obtient en faisant varier le niveau α (et donc ici la fonction quantile g_α, mais ceci est utilisé plus généralement) si les données nous amènent à rejeter pour une valeur de α elles amènent à rejeter pour toute valeur plus grande. Il est donc intéressant de connaitre la quantité (aléatoire, fonction des données) qui nous indique le plus petit niveau pour lequel les données rejettent. La $p-$value est donc un indice de signifiance de l'hypothèse nulle H_0. Plus la $p-$ value est grande, plus H_0 doit être acceptée. Réciproquement, évidemment plus elle est petite plus on a tendance à la rejeter.

Il est clair que le modèle linéaire est d'autant mieux adapté aux données que la variance expliquée est plus grande ou bien la variance résiduelle est plus faible c'est-à-dire que l'angle ω entre le vecteur centré $Y - \overline{Y}1_n$ (prédiction par une constante) et le vecteur $Y - \hat{Y}$ est plus proche de $\pm\pi/2$. De facon équivalente, on s'intéresse donc traditionnellement au sinus de cet angle.

$$sin^2\omega = R^2 = \frac{\|\hat{Y} - \bar{Y}1_n\|^2}{\|Y - \bar{Y}1_n\|^2}.$$

Il est facile de voir qu'on a la relation suivante entre R^2 et notre statistique de test F (d'où son nom)

$$F = \frac{n-p}{p-1} \quad \frac{R^2}{1-R^2}.$$

Le R^2 est une quantité qui se donne systématiquement dans les logiciels lorsqu'on fait une régression. Cependant il faut noter que lorsque la constante 1_n n'appartient pas au plan de régression, le R^2 défini comme précédemment n'a plus grande signification. On peut changer de définition et introduire $R^{2\prime}$ le cosinus de l'angle entre Y et son ajusté $\hat{Y}$.

$$cos^2\theta = \frac{{}^t\hat{Y}\hat{Y}}{{}^tyy} = 1 - \frac{{}^t\hat{\epsilon}\hat{\epsilon}}{{}^tYY}.$$

Cette quantité aussi permet de qualifier l'adéquation du modèle linéaire à nos données.

2.7.2 Etude de la validité du modèle : Tests non paramétrique sur les résidus

Plus haut nous avons construit des nouvelles variables $\eta_1, \ldots, \eta_{n-p}$ à partir des résidus on peut construire des nouvelles variables qui, elles sont i.i.d.$N(0, \sigma^2)$(et indépendantes de $\hat{\beta}$) :

Ces nouvelles variables (fonction des observations) peuvent nous servir à tester le modèle. On peut en effet tester l'hypothèse H_0 : les η_i sont i.i.d. $N(0, \sigma^2)$, contre H_1 : il existe $m \neq 0$ tel que les η_i sont i.i.d. $N(m, \sigma^2)$, qui correspondrait à l'oubli d'un centrage par exemple.

En général on a tendance à ne pas avoir d'idée sur la forme de ce qu'on pourrait avoir oublié dans le modèle on a alors recours à des tests de type non paramétriques.

On peut par exemple si σ^2 est connu, utiliser un test de Kolmogorov Smirnov. Si σ est inconnu, on peut 'standardiser' c'est à dire diviser les η_i par un estimateur bien choisi de σ. Le problème alors est que les η_i une fois standardisées ne sont plus i.i.d.... On peut aussi utiliser un test de signes ou de rangs ou de signes et rangs sur le η_i.

Souvent les logiciels prennent d'assez grandes libertés avec la théorie puisqu'ils proposent fréquemment un test de Kolmogorov Smirnov calculé directement sur les résidus $\hat{\varepsilon}$ standardisés ou fournissent des indices graphiques ($Q\times Q$ plot,...).

2.7.3 Significativité de chacune des variables explicatives

On s'intéresse à éliminer de l'étude toutes les variables non significatives pour le modèle proposé. Pour chaque variable explicative X^j, on veut effectuer le test

$$H_0 : \quad \beta_j = 0 \quad \text{contre} \quad H_1 : \quad \beta_j \neq 0$$

qui revient à tester

$H_0 :$ X^j est non significative contre $H_1 :$ X^j est significative .

Dans ce cas, le test étudié plus haut nous permet de construire la statistique

$$F = \frac{\|P_{W_1} Y\|^2/l}{\|\hat{\varepsilon}\|^2/(n-p)}$$

où ici $l = 1$. Prenons le cas (les autres s'en déduisent par permutation des colonnes) $j = p$. Il est facile de voir que, si on reprend l'orthonormalisation de Gram Schmidt détaillée au paragraphe 2.3.2 ainsi que le résultat de ce paragraphe, $W_1 = sp\{Z^p\}$, $P_{W_1} Y = \frac{\langle Y, Z^p\rangle}{\langle Z^p, Z^p\rangle} Z^p = \hat{\beta}_p Z^p$. De sorte que la statistique de test s'écrit :

$$F = \frac{\hat{\beta}_p^2 \|Z^p\|^2}{\|\hat{\varepsilon}\|^2/(n-p)}.$$

On peut soit calculer directement $\|Z^p\|^2$ soit remarquer que cette quantité doit nécessairement être l'inverse de la variance de $\hat{\beta}_p$ (divisée par σ^2), ce qu'on a aussi calculé au paragraphe 2.6.1 et vaut x_{pp} le p-ème élément de la diagonale de la matrice $({}^tXX)^{-1}$ (mais cela demande alors de l'avoir inversée exactement).

En remarquant qu'un loi $F(1, n-p)$ est le carré d'une loi de Student $T(n-p)$, on a tendance (ce qui est strictement équivalent) à utiliser comme statistique de test

$$T = \frac{\hat{\beta}_j}{\sqrt{\widehat{\sigma}^2 x_{jj}}}$$

où x_{jj} est le j-ième élément de la diagonale de la matrice $({}^tXX)^{-1}$. Sous l'hypothèse nulle H_0, T suit donc une loi de student à $n-p$ degrés de liberté. Pour tester la significativité du régresseur X^j au niveau α, on trouve donc le $\alpha-$quantile q_α de la loi t_{n-p} et on applique la règle de décision

– si $|T| > q_\alpha$, on refuse H_0 et X^j est significative,
– si $|T| < q_\alpha$, on accepte H_0 et X^j n'est pas significative.

Bien sur, on peut aussi utiliser la $p-$value pour prendre la décision.

2.8 Multi-colinéarité

Pour estimer les paramètres et leur variance, on a besoin de calculer l'inverse de la matrice $({}^tXX)$. Lorsque le déterminant de cette matrice est nul ou très proche de 0, on dit que le problème est **mal conditionné**. On est confronté à des estimateurs qui ont des grandes variances (donc peu précis) et il apparait souvent des problèmes de précision numérique. Il faut donc pouvoir diagnostiquer ces situations et proposer des solutions.

2.8.1 Diagnostics

La matrice de variance-covariance de l'estimateur des MCO s'écrit

$$V = \sigma^2({}^tXX)^{-1}$$

et on a montré dans le paragraphe précédent que chaque élément de la diagonale de cette matrice (qui est la variance des paramètres estimés) peut s'exprimer sous la forme suivante : prenons d'abord le dernier pour faire les calculs

$$\begin{aligned} V_{pp} &= \frac{1}{\|Z^p\|^2} \\ &= \frac{1}{\|X^p - P_{sp\{X^1,\ldots,X^{p-1}\}}X^p\|^2} \\ &= \frac{1}{\|X^p\|^2[1 - \frac{\|P_{sp\{X^1,\ldots,X^{p-1}\}}X^p\|^2}{\|X^p\|^2}]}. \end{aligned}$$

Soit encore

$$V_{jj} = \frac{1}{\|X^j\|^2(1 - R_j^2)}$$

où R_j^2 est le coefficient de détermination de la variable X^j sur celles qui restent (c'est le cosinus carré de l'angle entre X^j et la projection de X^j sur l'espace engendré par les autres variables $X^1, \ldots X^{j-1}, X^{j+1}, \ldots X^p$). Il est évident que plus X^j est linéairement proche de cet espace, plus R_j^2 est proche de 1 et plus V_{jj} est grand. Cette variance est minimum (c'est-à-dire l'estimateur est le plus précis) lorsque X^j est orthogonale aux autres variables. On appelle V_{jj} le **facteur d'inflation de la variance**.

En examinant la matrice des corrélations entre les variables, on peut détecter les variables très corrélées 2 à 2 mais pas les corrélations multiples. Il faut donc calculer effectivement les V_{jj} ou plutôt les **tolérances** $1 - R_j^2$.

Pour regarder les problèmes de colinéarité 2 à 2, on peut calculer **l'indice de conditionnement**

$$\kappa = \max(\lambda_j)/\min(\lambda_j),$$

où $\lambda_j, j = 1, \dots p$ sont les valeurs propres de la matrice des corrélations. En pratique si $\kappa < 100$, on considère qu'il n'y a pas de problème. Par contre, il faut s'inquiéter si $\kappa > 1000$. Cet indice donne une idée globale des problèmes de colinéarité mais pour savoir quelles variables posent problème, il faut calculer les facteurs d'inflation et les tolérances.

2.8.2 Modèles curvilinéaires

En cas de non validité de l'hypothèse de linéarité, il est intéressant de considérer des modèles polynomiaux

$$Y = \beta_1 + \dots \beta_p X^p + \dots c_{kl} X^k X^l + \dots d_j (X^j)^2 + \dots$$

qui sont appelés aussi **surfaces de réponse**. Ces modèles sont très simples à étudier : il suffit de rajouter les nouvelles variables produit des anciennes. Attention, ce type de modèles accroit les risques de colinarité : dans la pratique, il est rare de considérer des modèles autres que quadratiques.

2.9 Sélection de variables et Choix de modèles

La modélisation statistique couvre 3 objectifs

1. **description :** on veut explorer les liaisons entre Y et $X^1, \dots X^p$ pour p grand. Le but est de sélectionner un sous ensemble de variables explicatives dont le cardinal n'est pas trop grand. Attention, si n est petit et p grand, il est toujours possible de trouver un "bon" modèle : c'est l'effet *data mining*.
2. **explication :** on a des connaissances a priori et on veut valider ou invalider ces résultats théoriques. Le modèle exploratoire précédant permet de faire de l'inférence : tests et intervalles de confiance.
3. **prédiction :** On veut avoir de "bons" estimateurs (par rapport au critère de risque quadratique par exemple) afin de faire des prédictions correctes. On veut en général trouver des modèles *parcimonieux* (c'est-à-dire avec peu de variables explicatives). On préfère avoir des modèles avec des estimateurs légèrement biaisés pour avoir un bon compromis biais/variance. Ici, un "bon" modèle n'est plus celui qui explique le mieux (bon R^2 ou petite SCR) mais celui qui prédit le mieux.

Il existe beaucoup de critères permettant de choisir le modèle : AIC, BIC, erreur quadratique de prédiction .... Ils sont tous équivalents lorsqu'on fixe le nombre de variables p à sélectionner. Mais, par contre le choix du critère joue un rôle important lorsqu'on veut comparer 2 modèles utilisant un nombre différent de variables explicatives.

2.9.1 Statistique de Fisher :

On utilise ce critère pour comparer des suites de modèles emboités. Rappelons qu'on a aussi utilisé la statistique de Fisher dans le cadre explicatif pour tester la validité globale d'un modèle (test du R^2).

On a un modèle (gros) avec p variables noté M_1, un modèle (petit) avec q variables (choisies parmi les p utilisées dans le "gros" modèle) noté M_0. On calcule la statistique de Fisher où pour bien marquer la dépendance dans les variables nous écrirons $RSS0(q)$ et $RSS1(p)$ à la place de $RSS0$ et $RSS1$

$$F = \frac{(RSS0(q) - RSS1(p))/(p-q)}{RSS1(p)/(n-p)} = \frac{n-p}{p-q}[\frac{RSS0(q)}{RSS1(p)} - 1].$$

Si cette statistique est assez grande (supérieure à $f_\alpha(p-q, n-p)$) alors l'ajout des $p-q$ variables supplémentaires est justifié. Sinon, on peut se contenter du petit modèle à q variables. Plus exactement, cette statistique permet d'effectuer le test

$$H_0 : M_0 \text{ valide} \qquad \text{contre} \qquad H_1 : M_1 \text{ valide}$$

soit

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0 \text{ contre } H_1 : \exists j \in \{q+1, \ldots, p\}, \quad \beta_j \neq 0.$$

2.9.2 Critères de choix : AIC, BIC, C_p

Il y a un problème avec ce type de test, c'est qu'on ne contrôle vraiment son erreur que si on le pratique une fois pour un choix bien précis de variable. Il est clair que si on fait plusieurs tests les uns après les autres pour choisir les variables, le calcul du niveau devient très vite fastidieux. On ne procède pas de cette façon mais on conserve l'idée de regarder les fluctuations de la statistique. Supposons que nos variables soient ordonnées et que l'on se pose la question d'en rajouter de plus en plus. Dans ce cas, on ne procède pas avec un test mais on conserve l'idée de regarder les fluctuations de la statistique.

$$q \mapsto \frac{(n-p)RSS0(q)}{RSS1(p)}$$

qui représente bien l'erreur que l'on fait en prédisant le modèle si on s'arrête aux q premières variables normalisée par l'erreur faite avec toutes les variables possibles. Evidemment, à mesure que 'lon augmente le nombre q de variables explicatives, cette statistique se rapproche de 1. Donc cela ne nous donne pas un critère de choix : on prédit d'autant mieux qu'on a plus de variables explicatives.

Pour remédier à ce problèmes plusieurs critères sont proposés dans la littérature, qui consistent à pénaliser le nombre de variables explicatives.

Citons parmi eux

$$BIC(q) = \frac{(n-p)RSS0(q)}{RSS1(p)} + [\log n]q \quad (Schwarz\,'76) \tag{2.4}$$

$$AIC(q) = \frac{(n-p)RSS0(q)}{RSS1(p)} + q. \quad (Akaike\,'70,'73) \tag{2.5}$$

$$C_p(q) = \frac{(n-p)RSS0(q)}{RSS1(p)} + 2q. \quad (Mallows\,'73) \tag{2.6}$$

2.9.3 Algorithmes de sélection

Avec p variables explicatives, on 2^p choix de modèles possibles. Si p est grand, il n'est pas raisonnable d'explorer tous les modèles pour trouver le meilleur. Il existe 3 types d'algorithmes :

1. Pas à pas :
   - **forward :** On commence avec une variable et à chaque pas, on en ajoute une : celle qui apporte le plus pour le critère de la statistique de Fisher. On s'arrête soit lorsqu'il n'y a plus de variable, soit quand aucune variable n'apporte quelque chose ou en applicant un critère de type AIC ou BIC.
   - **backward :** On fait la même chose mais en démarrant du modèle complet. On élimine la variable qui apporte le moins par rapport au critère de Fisher. On s'arrête lorsque les variables restantes donnent toutes un critère satisfaisant (pour un α fixé à l'avance).
   - **stepwise :** Après chaque sélection de modèle donnée par la méthode "forward", on enlève les variables qui deviennent inutiles du fait de l'ajout de nouvelles variables.
2. Par échange :
   - **maximisation du R^2 :** On travaille avec un nombre q fixé de variables explicatives du modèle. On cherche alors une nouvelle variable qui maximise l'accroissement du R^2. Puis, on cherche avec quelle variable présente dans le modèle l'échanger de facon à rester avec q variables. On recommence tant que le R^2 croît.
   - **minimisation du R^2 :** Idem que précédemment mais on sélectionne la variable qui minimise l'accroissement du R^2. On explore alors plus de modèles et on a plus de chance de tomber sur un meilleur optimum.
3. Global : L'algorithme de Furnival et Wilson est utilisé pour comparer tous les modèles possibles en optimisant le R^2, ou un critère de type C_p AIC ou BIC. L'algorithme parcourt un arbre, évite les sous branches dont on sait a priori qu'elles ne sont pas compétitives. En général, les logiciels donnent le meilleur modèle pour chaque q. Mais ceci n'est possible que pour un nombre raisonnable de variables explicatives.

2.10 Théorèmes de Student et de Cochran

Théorème 2 (Student) *Soit $X_1, \ldots, X_n$, des variables indépendantes identiquement distribuées (notation i.i.d.) de loi commune $N(m, \sigma^2)$. Alors,*

1. $\bar{X}_n = \sum_{i=1}^{n} X_i$ *suit une loi* $N(m, \sigma^2/n)$.
2. $R_n = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ *suit une loi* $\sigma^2 \chi(n-1)$.
3. $\bar{X}_n$ *et* R_n *sont indépendants.*

4. *Si S_n désigne la variable $\sqrt{\frac{R_n}{n-1}}$, alors $T_n = \frac{\sqrt{n}(\bar{X}_n - m)}{S_n}$ suit une loi de Student T(n-1).*

Démonstration du Thérème de Student

- 1 est évident.
- Les quantités que nous étudions sont homogènes. Par le changement de variables $X_i' = (X_i - m)/\sigma$, on se ramène au cas où $m = 0,\ \sigma^2 = 1$.
- Notons qu'on a la relation suivante :

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + n\bar{X_n}^2 = \sum_{i=1}^{n} X_i^2. \tag{2.7}$$

On considère une matrice orthogonale M telle que sa première ligne est $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$. Soit $Z = MX$ où $X = (X_1, \ldots, X_n)^*$. Puisque M est orthogonale, Z est un vecteur gaussien standard de $\mathbb{R}^n$, et $Z_1 = \sqrt{n}\bar{X}_n$ est indépendant de $(Z_2, \ldots, Z_n)$. Par ailleurs, toujours parce que M est orthogonale,

$$\|MX\|^2 = \|X\|^2 = \sum_{i=1}^{n} X_i^2 = (\sqrt{n}\bar{X}_n)^2 + \sum_{i=2}^{n} Z_i^2.$$

On en déduit que $\sum_{i=2}^{n} Z_i^2 = \sum_{i=1}^{n} X_i^2 - (\sqrt{n}\bar{X}_n)^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ (en utilisant (2.7)) est indépendant de $\bar{X}_n$ et suit un $\chi^2(n-1)$.

Théorème 3 (COCHRAN) *Soit $X \sim N(\xi, I_n)$*

1. *Soit $P_1, P_2, \ldots, P_k$, k matrices $n \times n$ autoadjointes, verifiant*

$$I_n = \sum_{i=1}^{d} P_i, \quad et \quad \sum_{i=1}^{d} rangP_i \leq n.$$

Alors les matrices P_i sont des projecteurs ($P_i^2 = P_i$) et les variables P_iX sont des Gaussiennes mutuellement indépendantes de loi $N(P_i\xi, P_i)$.

2. *Soit $Q_1, Q_2, \ldots Q_k$, k formes quadratiques sur $\mathbb{R}^n$ verifiant :*

$$\forall x \in \mathbb{R}^n, \quad \|x\|^2 = \sum_{i=1}^{d} Q_i(x) \quad et \quad \sum_{i=1}^{d} rangP_i \leq n.$$

Alors les variables Q_iX sont mutuellement indépendantes de loi $\chi'^2(Q_i\xi, rangQ_i)$.

Démonstration du Théorème : La démonstration repose sur un lemme de pure algébre linéaire :

Lemme 1 *Soit $P_1, P_2, \ldots, P_k$, k matrices $n \times n$, vÈrifiant*

$$I_n = \sum_{i=1}^{d} P_i, \quad et \quad P_i = P_i^*$$

On a alors l'equivalence entre :

1. $\sum_{i=1}^{d}$ *rang* $P_i \leq n$.
2. $\forall i \neq j \quad P_i P_j = 0$
3. $\forall i \quad P_i^2 = P_i$

Preuve du Lemme : Remarquons que 1 signifie : $\forall x \in \mathbb{R}^n, \quad x$ s'écrit de manière unique sous la forme $\sum_{i=1}^{k} u_i; \quad u_i \in P_i(\mathbb{R}^n)$.

1. $2 \Rightarrow 3$ $P_i = P_i(\sum_j P_j) = \sum_j P_i P_j = P_i^2$
2. $3 \Rightarrow 2$ On a

$$\forall x \in \mathbb{R}^n, \quad \|x\|^2 = \langle x, x\rangle = \langle x, \sum_j P_j x\rangle = \langle x, \sum_j P_j^2 x\rangle = \sum_j \|P_j x\|^2.$$

Appliquons cette relation à $P_i x$:

$$\forall x \in \mathbb{R}^n, \quad \|P_i x\|^2 = \sum_j \|P_j P_i x\|^2 = \|P_i x\|^2 + \sum_{j \neq i} \|P_j P_i x\|^2.$$

Donc $j \neq i \Rightarrow P_j P_i = 0$

3. 3&2 $\Rightarrow 1$ Soit $x = \sum_i P_i y_i$. On a donc :

$$P_j x = \sum_i P_j P_i y_i = P_j^2 y_j = P_j y_j.$$

D'o˘ l'Écriture unique $x = \sum_i P_i x$.

4. $1 \Rightarrow 3\&2$ $P_j = (\sum_i P_i) P_j = \sum_i P_i P_j$. On en déduit ;

$$\forall x \in \mathbb{R}^n, \quad P_j(x - P_j x) = \sum_{i \neq j} P_i P_j x.$$

L'unicité de la representation implique le resultat.

Démonstration du Théorème, (fin)

1. C'est une conséquence du fait que pour des vecteurs gaussiens orthogonalité signifie indépendance.
2. Soit $P_j = P_j^*$ la matrice definissant la forme quadratique Q_j : $\forall x \in \mathbb{R}^n \quad Q_j(x) = x^* P_j x$. Par polarisation de la relation $\forall x \in \mathbb{R}^n, \quad \|x\|^2 = \sum_{i=1}^{d} Q_i(x)$, on obtient :

$$\forall x, y \in \mathbb{R}^n, \quad \langle x, y\rangle = \sum_j \langle x, P_j y\rangle$$

ce qui implique $I_n = \sum_j P_j$. Le point 2 du théorème est donc une conséquence du point 1 et de la proposition 3 suivante.

Proposition 3 .

1. *Si P est une matrice de projection (i.e. $P = P^* = P^2$), et si $W \sim N(\xi, P)$, avec $P(\xi) = \xi$, alors $\|W\|^2 \sim \chi'^2(rang\ (P), \|\xi\|^2)$*

2. *Si P est une matrice de projection (i.e. $P = P^* = P^2$), et si $X \sim N(\xi, I_n)$ alors, $\|PX\|^2 \sim \chi'^2(rang\ (P), \|P(\xi)\|^2)$.*

Démonstration de la Proposition :

1. En effet , on peut écrire, au moyen de la matrice R orthogonale, $P = RDR^*$ où D est une matrice diagonale dont les $d = (rang(P))$ premiers coefficients sont égaux à 1, les autres à 0. Soit $Z = R^*W$. On a $W = RZ$, et $Z \sim N(\eta, D), \quad R^*\xi = \eta$. Comme $\xi = RDR^*\xi$, on a $\eta = D\eta$.
   Donc les $n-d$ dernières composantes de Z sont nulles, et les d premières,suivent des lois normales $N(\eta_i, 1)$ indépendantes. De plus $\sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{d} \eta_i^2$.
   Comme $\|W\|^2 = \|Z\|^2, \quad \|W\|^2 \sim \chi'^2(d, \|\xi\|^2)$.
2. On remarque $PX \sim N(P\xi, P)$.