(*Cours 1: fin*)

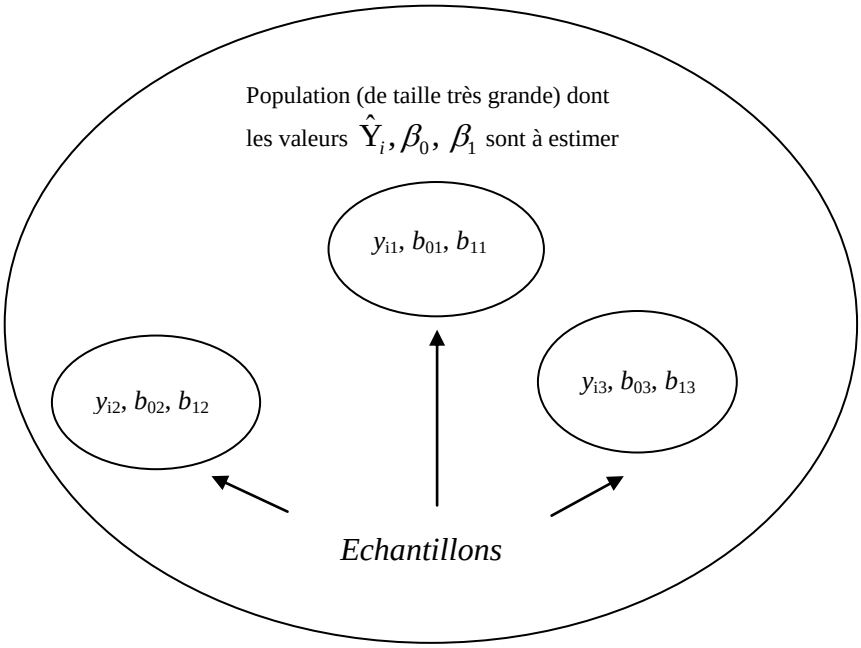
I . INFERENCE SUR LES PARAMETRES DU MODELE DE REGRESSION SIMPLE.

L'inférence statistique présente en régression plusieurs aspects. Par exemple :

- Déterminer l'intervalle de confiance sur β_1
- Tester si la régression est significative
- Déterminer la marge d'erreur dans l'estimation de la moyenne conditionnelle

I.1. Propriétés des estimateurs b0 et b1. Chaque estimateur est une combinaison linéaire des observations Yi. Ce sont des estimateurs linéaires. Ils sont des estimateurs sans biais, c.-à-d. : l'espérance mathématique de b0 et b1 est respectivement β_0 et β_1 :

$$E(b0) = \beta_0 \; ; \; E(b1) = \beta_1$$



Théorème.

Parmi tous les estimateurs linéaires non biaisés de β_0 β_1, b_0 et b_1 sont ceux qui présentent une variance minimale.

On en conclut qu'ils sont les meilleurs estimateurs à notre disposition.

Pour déterminer un intervalle de confiance sur un paramètre ou bien exécuter un test statistique, il nous faut connaître la distribution d'échantillonnage de l'estimateur étudié : forme, moyenne et variance. Pour cela nous n'avons à notre disposition les Yi (les données), les e_i et les propriétés vues ci-dessus.

Distribution d'échantillonnage de b_1

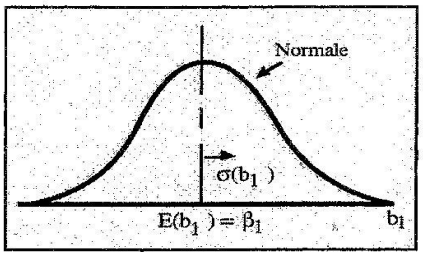
La distribution d'échantillonnage de l'estimateur b_1 est une distribution **normale** de moyenne $E(b_1) = \beta_1$ et de variance

$$Var(b_1) = \sigma^2(b_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

où $\sigma^2 = Var(Y_i) = Var(\varepsilon_i)$.

Les fluctuations de l'écart réduit $Z = \frac{b_1 - \beta_1}{\sigma(b_1)}$ suivent la loi normale centrée réduite où

$$\sigma(b_1) = \sqrt{Var(b_1)} = \frac{\sigma}{\sqrt{\sum (X_i - \overline{X})^2}}$$



Estimation de la variance de b_1

La variance $\sigma^2(b_1)$ est habituellement inconnue du fait que σ^2 est inconnue. L'estimation de σ^2 s'obtient de s^2, la variance des résidus et l'esimation de la variance de b_1, que nous notons $s^2(b_1)$, s'obtient alors de

$$s^2(b_1) = \frac{s^2}{\sum (X_i - \overline{X})^2}$$

L'écart-type de b_1 s'écrit $s(b_1) = \frac{s}{\sqrt{\sum (X_i - \overline{X})^2}}$

Fluctuations de l'écart réduit: petit échantillon

Dans la pratique, on ne connaît pas σ^2; on doit donc l'estimer avec s^2. De plus, il est fréquent que l'on doit se contenter d'un petit échantillon, disons moins d'une trentaine d'observations. Quelle est alors la distribution de l'écart réduit $\frac{b_1 - \beta_1}{s(b_1)}$?

Écart réduit $\frac{b_1 - \beta_1}{s(b_1)}$: conditions d'application et distribution

Sous les conditions suivantes:

i) b_1 est distribué normalement,
ii) la taille d'échantillon est petite, $n - 2 < 30$,
iii) la variance $\sigma^2(b_1)$ est inconnue mais est estimée par $s^2(b_1)$,

alors les fluctuations de l'écart réduit $t = \frac{b_1 - \beta_1}{s(b_1)}$ sont celles de la loi de Student avec $\nu = n - 2$ degrés de liberté.

I.2. Estimation de β1 par intervalle de confiance. A moins d'être en présence d'un grand échantillon, auquel cas on utilise la loi normale centrée réduite pour établir l'intervalle (avec $s^2 \cong \sigma^2$), on utilise la loi de Student.

A partir d'un échantillon aléatoire de petite taille ($n-2 < 30$) d'observations Y_i distribués selon une loi normale de moyenne $E(Y_i) = \beta_0 + \beta_1 X_i$ et de variance σ^2 inconnue, on prend comme estimation ponctuelle b1 de β1 pour établir un intervalle ayant un niveau de confiance $100(1-\alpha)\%$ de contenir la véritable valeur de β1, comme suit :

$$b_1 - t_{\alpha/2;\nu} \cdot s(b_1) \le \beta_1 \le b_1 + t_{\alpha/2;\nu} \cdot s(b_1)$$

où $\nu = n - 2$ degrés de libertés et $s(b_1) = \frac{s}{\sqrt{\sum (X_i - \overline{X})^2}}$

et $t_{\alpha/2;\nu}$ est la valeur tabulée de la distribution de Student telle que la probabilité que t soit comprise entre $-t_{\alpha/2;\nu}$ *et* $+t_{\alpha/2;\nu}$ est $(1-\alpha)$

Exemple

Dans une étude sur l'analyse des propriétés d'un sol dans la région au sud du fleuve St-Laurent, on a obtenu les résultats suivants pour une étude de régression entre la teneur (Y) en fer (ppm) et le pourcentage (X) de matières organiques du sol pour 40 observations:

$$b_1 = 17{,}847, \quad s^2 = 3232{,}05, \quad \sum (X_i - \overline{X})^2 = 22{,}931.$$

Déterminons un intervalle de confiance pour β_1 avec un niveau de confiance de 95%; β_1 représente ici la variation de la teneur moyenne en fer pour une variation unitaire du pourcentage de matières organiques.

On a $b_1 = 17{,}847$, $n = 40$, $s^2 = 3232{,}05$,

d'où $$s^2(b_1) = \frac{s^2}{\sum (X_i - \overline{X})^2} = \frac{3232{,}05}{22{,}931} = 140{,}9467$$

et $$s(b_1) = \sqrt{140{,}9467} = 11{,}872.$$

Avec $\nu = n-2 = 40-2 = 38$, on obtient (de la table de Student) $t_{0{,}025;38} = 2{,}0244$.

En substituant dans l'intervalle $b_1 - t_{\alpha/2;\nu} \cdot s(b_1) \le \beta_1 \le b_1 + t_{\alpha/2;\nu} \cdot s(b_1)$, on obtient :

$$-6{,}187 \le \beta_1 \le 41{,}881$$

Quelle remarque faites-vous ? Comment peut-on tester l'hypothèse H0 : $\beta 1 \approx 0$ *?*

Alternative : la statistique qui convient pour le test est b_1. Si H0 est vraie, alors l'écart réduit **t** est distribué selon la loi de Student avec (n-2) ddl :

$$t = \frac{b_1 - \beta_1}{s(b_1)} = \frac{b_1}{s / \sqrt{\sum (X_i - \overline{X})^2}}.$$

Règle : rejeter H0 si $t > t_{\alpha/2;n-2}$ ou bien si $t < -t_{\alpha/2;n-2}$

I.3 Inférence concernant β_0. Il est moins fréquent de tester statistiquement le paramètre β_0 (R le fait systématiquement !). Tout comme dans le cas précédent on calcule les fluctuations de l'écart réduit $t = \frac{b_0 - \beta_0}{s(b_0)}$ avec $s(b_0) = s \cdot \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right]^{1/2}$. La règle est identique à la précédente.

I.4 Intervalle de confiance pour E(Yh) à X = Xh. Il s'agit d'établir maintenant un intervalle de confiance autour d'une estimation ponctuelle de **E(Yh)** par la moyenne $\hat{Y}_h = b_0 + b_1 X_h$ au niveau de confiance 100(1-α)% de contenir la vraie valeur **E(Yh) :**

$$\hat{Y}_h - t_{\alpha/2;\nu} \cdot s(\hat{Y}_h) \le \beta_1 \le \hat{Y}_h + t_{\alpha/2;\nu} \cdot s(\hat{Y}_h)$$

Pour les petits échantillons, on peut affirmer que $t = \frac{\hat{Y}_h - E(\hat{Y}_h)}{s(\hat{Y}_h)}$ est distribuée selon la loi de Student avec (n-2) ddl. On en déduit que l'intervalle de confiance s'établit en calculant :

$$s(\hat{Y}_h) = s \cdot \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right]^{1/2} \text{ et } \hat{Y}_h = b_0 + b_1 X_h$$

Remarque : Cet intervalle de confiance a une amplitude variable puisqu'il dépend de $s(\hat{Y}_h)$ et que ce dernier varie selon la position relative de X_h par rapport à $\overline{X}$. (remarquer sa valeur particulière pour $X_h = 0$).

I.5 Intervalle de confiance pour E(Yh) à X = Xh où Xh est une nouvelle observation.

Intervalle de confiance $\hat{Y}_h - t_{\alpha/2;\nu} \cdot s(d_h) \leq \beta_1 \leq \hat{Y}_h + t_{\alpha/2;\nu} \cdot s(d_h)$

Où $s(d_h) = s \cdot \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right]^{1/2}$ et $\hat{Y}_h = b_0 + b_1 X_h$

Cours 2

II. LA REGRESSION MULTIPLE.

Nous avons traité précédemment d'un modèle dans lequel une variable est dépendante d'une unique variable explicative :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

La démarche consistait à estimer les paramètres du modèle, à décomposer la variance de la variable dépendante en diverses sources de variation sous forme d'un tableau d'analyse de variance, à tester si β_1 est significativement différent de 0, à estimer les intervalles de confiance autours des estimateurs et autour de nouvelles valeurs prédites. Nous considérons maintenant le cas où une variable est dépendante *de plusieurs variables* explicatives.

⚠ *Le terme de linéaire s'applique aux paramètres du modèle et non aux variables. La régression sera donc valide même si le modèle comporte des termes non linéaires en X (puissance, produit de variables…).*

La forme générale du modèle de régression multiple s'écrit ainsi :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

Hypothèses

- Le terme ε_i est une variable aléatoire de moyenne = 0 et de variance constante = $\text{Var}(\varepsilon_i) = \sigma^2$
- Il n'existe aucune corrélation entre les ε_i, i.e. $\text{Cov}(\varepsilon_i\ \varepsilon_j) = 0$, pour tout i et tout j ($i \neq j$).
- Les variables explicatives sont des grandeurs certaines (as usual…)

On en déduit que les observations *Yi* sont distribuées normalement et indépendamment avec :

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$$

$$\text{Et } \text{Var}(Y_i) = \sigma^2$$

Contraintes :

- Il n'existe aucune colinéarité entre variables explicatives
- Le nombre de données excède le nombre de paramètres à estimer : n > k + 1.

II.1 Interprétation des paramètres du modèle

β_0 représente le niveau moyen des Yi lorsque chacune des variables explicatives est nulle. On peut lui donner une signification particulière.

Les β_i représentent les changements subis par E(Yi) attribuables à un changement unitaire de la ième variable, les autres restant inchangées. Ainsi : $\beta_j = \dfrac{\partial E(Y_i)}{\partial X_{ij}}$.

Par exemple : $\beta_3 = \dfrac{\partial E(Y_i)}{\partial X_{i3}} \Big| X_{i1}, X_{i2}, X_{i4}$ fixes , représente le changement subi par Yi correspondant à une variation unitaire de la variable 3, les autres variables restant inchangées. NB : un nombre de 5 données minimal est requis.

II.2 Détermination de l'équation de régression multiple

L'estimateur de E(Yi) est $\hat{Y}_i = b_0 + b_1 X_{i1} + b_{i2} X_{i2} + ... + b_k X_{ik}$ qui est l'équation de régression. On applique la méthode des moindres carrés qui conciste à minimiser la somme des carrés résiduelle soit :

$$\text{Min}(\sum_{i=1}^{n} e_i^2) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - ... - b_k X_{ik})^2$$

Comme précédemment, on a recours aux dérivées partielles résoudre ce problème.

$$\frac{\partial(\sum_{i=1}^{n} e_i^2)}{\partial b_0}, \frac{\partial(\sum_{i=1}^{n} e_i^2)}{\partial b_1}, \frac{\partial(\sum_{i=1}^{n} e_i^2)}{\partial b_2}, \frac{\partial(\sum_{i=1}^{n} e_i^2)}{\partial b_3}, ..., \frac{\partial(\sum_{i=1}^{n} e_i^2)}{\partial b_k}.$$

Annulant ces dérivées, et après quelques manipulations algébriques, on obtient le système de (k+1) équations que l'on veut résoudre pour $b_0, b_1, b_2, ..., b_k$.

$$nb_0 + b_1\sum X_{i1} + b_2\sum X_{i2} + b_3\sum X_{i3} + ... + b_k\sum X_{ik} = \sum Y_i$$

$$b_0\sum X_{i1} + b_1\sum X_{i1}^2 + b_2\sum X_{i1}X_{i2} + b_3\sum X_{i1}X_{i3} +...+ b_k\sum X_{i1}X_{ik} = \sum X_{i1}Y_i$$

$$b_0\sum X_{i2} + b_1\sum X_{i2}X_{i1} + b_2\sum X_{i2}^2 + b_3\sum X_{i2}X_{i3} +...+ b_k\sum X_{i2}X_{ik} = \sum X_{i2}Y_i$$

$$\vdots$$

$$b_0\sum X_{ik} + b_1\sum X_{ik}X_{i1} + b_2\sum X_{ik}X_{i2} + b_3\sum X_{ik}X_{i3} +...+ b_k\sum X_{ik}^2 = \sum X_{ik}Y_i.$$

La résolution d'un tel système est facilitée en utilisant l'approche matricielle.

Forme matricielle du modèle de régression multiple

Dans le cas d'un modèle de régression multiple comportant k variables explicatives, la relation s'écrit:

$$Y_i = \beta_0X_0 + \beta_1X_{i1} + \beta_2X_{i2} + ... + \beta_jX_{ij} + ... + \beta_kX_{ik} + \varepsilon_i, \quad i = 1,..., n$$

où X_0 est une variable utilitaire, $X_0 = 1$.

Identifions par le vecteur **Y** (vecteur colonne de dimensions $n \times 1$), les n observations associées à la variable dépendante:

$x_0 = 1$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2k} \\ & & & \vdots & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times (k+1)}{\mathbf{X}} \cdot \underset{(k+1) \times 1}{\beta} + \underset{n \times 1}{\varepsilon}$$

Forme matricielle du modèle de régression

Pour obtenir les quantités $b_0, b_1, b_2, ..., b_k$, l'on doit résoudre le système d'équations obtenues précédemment.

Les sommations qui multiplient les coefficients de régression dans ce système d'équations se représentent sous forme matricielle à l'aide de la matrice $\mathbf{X'X}$ de dimensions (k+1) x (k+1):

$$\mathbf{X'X} = \begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} & \cdots & \sum X_{ik} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \cdots & \sum X_{i1}X_{ik} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ik} \\ & & \vdots & & \\ \sum X_{ik} & \sum X_{ik}X_{i1} & \sum X_{ik}X_{i2} & \cdots & \sum X_{ik}^2 \end{bmatrix}$$

Matrice X'X

Elle correspond au produit de la matrice $\mathbf{X'}$ par la matrice $\mathbf{X}$ où $\mathbf{X'}$ représente la transposée de la matrice $\mathbf{X}$. Cette matrice est obtenue de $\mathbf{X}$ en interchangeant les lignes et les colonnes de sorte que les lignes de $\mathbf{X}$ deviennent les colonnes de la matrice transposée:

$$\mathbf{X'} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ & \vdots & & \\ X_{1k} & X_{2k} & \cdots & X_{nk} \end{bmatrix}$$

Matrice de dimensions (k+1) x n

Le second membre des équations s'écrit sous forme matricielle:

$$\mathbf{X'Y} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1} Y_i \\ \sum X_{i2} Y_i \\ \vdots \\ \sum X_{ik} Y_i \end{bmatrix}$$

Vecteur-colonne de dimensions (k+1) x 1

La forme matricielle des coefficients de régression s'écrit:

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

Vecteur-colonne de dimensions (k+1) x 1

On obtient alors pour le système d'équations à résoudre, l'équation matricielle suivante:

$$(\mathbf{X'X}) \cdot \mathbf{b} = \mathbf{X'Y}.$$

Forme matricielle du système d'équations

Résolution du système d'équations normales

La résolution de ce système d'équations s'obtient à l'aide de l'inverse de la matrice **X'X**. En prémultipliant les deux membres de l'équation matricielle par $(\mathbf{X'X})^{-1}$, on obtient:

$$(\mathbf{X'X})^{-1}\ \mathbf{X'X}\ \mathbf{b} = (\mathbf{X'X})^{-1}\ \mathbf{X'Y}.$$

Puisque $(\mathbf{X'X})^{-1}\ (\mathbf{X'X}) = \mathbf{I}$, la matrice-unité de même dimension que $(\mathbf{X'X})^{-1}$, alors la solution au système d'équations est

$$\mathbf{b} = (\mathbf{X'X})^{-1}\ \mathbf{X'Y}.$$

Détermination des coefficients de régression

Le calcul de la matrice inverse s'obtient à l'aide de:

$$(\mathbf{X'X})^{-1} = \frac{\text{adj}(\mathbf{X'X})}{|\mathbf{X'X}|}$$

où adj (**X'X**) est la matrice adjointe et $|\mathbf{X'X}|$, le déterminant de la matrice **X'X**. L'inverse existe en autant que $|\mathbf{X'X}| \neq 0$.

L'équation de régression multiple $\hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2} + \ldots + b_kX_{ik}$ s'écrit sous forme matricielle: $\hat{\mathbf{Y}} = \mathbf{Xb}$ où $\hat{\mathbf{Y}}$ est un vecteur-colonne de dimensions $n \times 1$.

NB : un exemple d'inversion de matrice est disponible ici : http://www.matheureka.net/Q119.htm

Applications

```
# Exemple de Régression Multiple
# downloader et sauver ours.txt dans votre répertoire
# sauvegarder ours.txt dans un dataframe au moyen de <read.table> (cf.
# Cours I)
# Les séparateurs du document sont des tabulations, la première ligne
# contient les noms des variables.

fit <- lm(y ~ x1 + x2 + x3, data = mydata)  #établissez votre modèle
summary(fit)                                # montre les résultats

# autres fonctions utiles
coefficients(fit)             # Les Coefficients
confint(fit, level=0.95)  # Intervalles de confiance des paramètres
fitted(fit)               # valeurs prédites
residuals(fit)            # et les résidus
anova(fit)                # Table anova
vcov(fit)                 # Matrice de covariance des paramètres du modèle
influence(fit)            # Diagnostic…
```

Quelques éléments pour l'héteroscédasticité, la normalité, ainsi que les observations.

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # Non obligatoire (4 graphes dans une fenêtre)
plot(fit)                      # Affichage

# Pour Comparer des modèles
# Comparer les modèles : Masse ~ Age + Hauteur + Longueur
# et Masse ~ Age + Hauteur

fit1 <- lm(y ~ x1 + x2 + x3 + x4, data = mydata)
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)
```

Ce qui conduit à se poser la question de la sélection des variables à partir, par exemple d'un grand ensemble. C'est un sujet délicat. En général on utilise des méthodes dites Stepwise. On introduit une à une les variables en testant si E(Yi) est significativement modifiée par l'introduction de cette variable supplémentaire. Pour cela R dispose de fonctions utiles.

```
# Régression par étape
library(MASS)                                  # un package à charger
fit <- lm(y~x1+x2+x3,data=mydata)      # le modèle de régression
step <- stepAIC(fit, direction="both") # une méthode stepwize
step$anova                                     # les résultats

# Autres outils
libray(car)
outlierTest(fit)
qqPlot(fit)
leveragePlots(fit)
```