« Cours Statistique et logiciel R »

Rémy Drouilhet [(1)], Adeline Leclercq-Samson [(1)], Frédérique Letué [(1)], Laurence Viry [(2)]

[(1)] Laboratoire Jean Kuntzmann, Dép. Probabilites et Statistique,

[(2)] Laboratoire Jean Kuntzmann, Dép. Modele et Algorithmes Déterministes

mars-mai 2016

Plan de la présentation

Qu'est ce que le modèle linéaire ?

- Modèle probabiliste simple, permettant de décrire et d'étudier la relation qui peut exister entre deux ou plusieurs variables.

- Modèle de base pour modéliser et analyser les variations d'une variable aléatoire Y quantitative en fonction de variables/facteurs qualitatifs et/ou quantitatifs

- Partant de n observations $(x_{i,1}, \cdots, x_{i,p}, y_i)_{i=1,\cdots,n}$, **construire un modèle du type**

$$Y = f(X_1, X_2, \cdots, X_p) + \epsilon$$

avec f linéaire, pour modéliser
  - la variabilité de la **variable à expliquer** Y
  - par celle d'une ($p = 1$) ou d'un groupe de **variables explicatives** ($p > 1$)

Quelques situations classiques d'emploi du modèle linéaire

Essentiellement deux cadres

1. **la régression linéaire** (simple ou multiple) pour modéliser la relation entre une variable aléatoire Y quantitative et une ou plusieurs variables quantitatives non aléatoires $x_1 \cdots, x_p$.

2. **l'analyse de la variance** pour apprécier l'effet de variables qualitatives (appelées facteurs) sur une variable quantitative Y
   - problème de comparaison de groupes

Combinaison des 2 points de vue dans **l'analyse de covariance** : comparer des régressions dans plusieurs groupes.

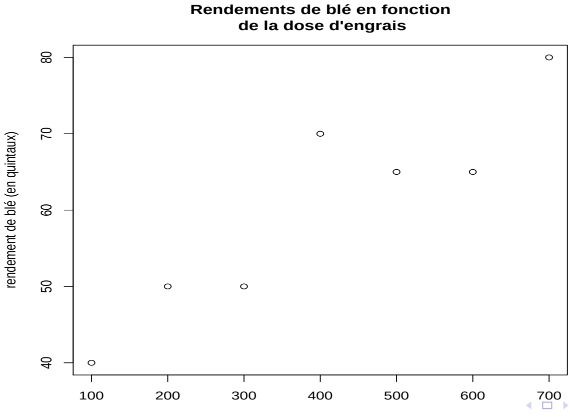
Exemple de régression linéaire simple

Les données et le problème

- Sur un échantillon de n=7 parcelles expérimentales choisies au hasard
- Etude du rendement en blé Y (en quintaux) en fonction de la quantité d'engrais X (en kg)
- Y : la **variable à expliquer** (ou la réponse),
- X : la **variable explicative** (ou facteur)
- Les agronomes pensent qu'il existe une relation linéaire entre le rendement de blé de parcelles d'une région donnée et la quantité d'engrais utilisée dans les parcelles

Les données

- Données : (xi, yi) $i = 1, \ldots, 7$
- x_i la quantité d'engrais utilisée sur la i^{eme} parcelle
- y_i le rendement de blé mesuré sur la i^{eme} parcelle

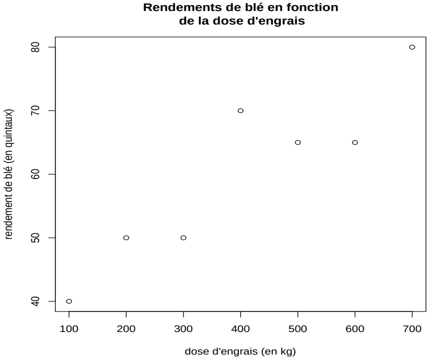


Objectifs

- **décrire** et **modéliser** la relation entre rendement et dose d'engrais
- **prévoir**, à partir du modèle, le rendement de blé que l'on obtiendrait en mettant une dose d'engrais de 450 kg, de 800 kg ?

Description des données

- Nuage de points



- $r(x, y) = 0.92$

Ajustement par la méthode des moindres carrés

- équation de la droite des **moindres carrés**

$$y = ax + b$$

a et b obtenus en minimisant la somme des carrés des erreurs

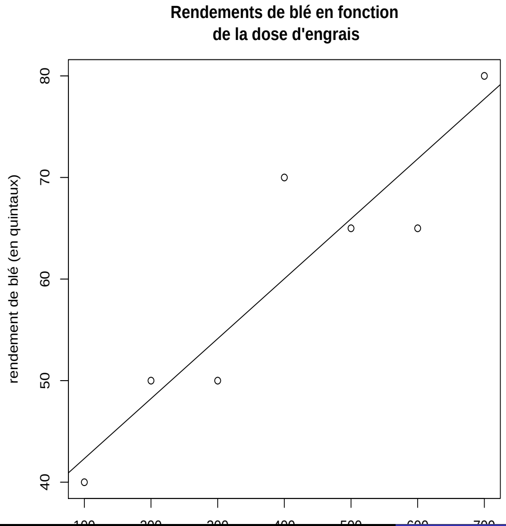
$$J(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

- minimisation de $J(a,b)$ en a et b conduit à

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov_{emp}(x,y)}{Var_{emp}(x)}$$

$$b = \bar{y} - a\bar{x}$$

Ajustement graphique



Droite des moindres carrés
$a = 0.059$ et $b = 36.429$

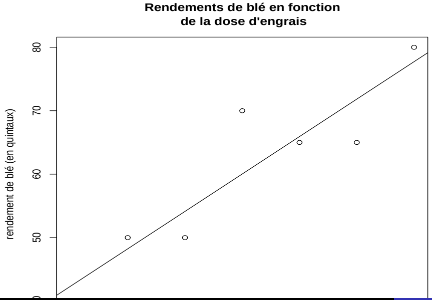
Qualité de l'ajustement
$R^2 = 0.845$

Prévision avec la droite des moindres carrés

- ajustement de bonne qualité
- prévision possible du rendement associé à une dose d'engrais x_0 avec la droite des moindres carrés

$$\hat{y}_0 \;=\; a\,x_0 + b$$

- pour $x_0 = 450\,kg$, on prévoit un rdt en blé de $\hat{y}_0 = \ldots\ldots\ldots$ quintaux
- pour $x_0 = 800\,kg$, on prévoit un rdt en blé de $\hat{y}_0 = \ldots\ldots\ldots$ quintaux



Les questions

- Dans un objectif de modélisation, quelle confiance accorder aux valeurs de a et b "déterminées" à partir de l'échantillon (x_i, y_i) de taille $n = 7$?
- Dans un objectif de prévision, comment apprécier la qualité de prévision et quelle confiance accordée à la valeur prédite ?

Réponse à partir de la théorie des probabilités

- Introduire un **modèle probabiliste** linéaire
- Modélisation du mécanisme de génération des données
- Considérer les **données comme la réalisation d'un échantillon de variables aléatoires** obéissant à certaines lois décrites par le modèle linéaire
- Etude "théorique" de ce modèle et validation
- Considérer les coefficients de la droite des moindres carrés comme les estimations des paramètres du modèle linéaire
- Obtenir des **résultats d'estimation ponctuelle, par intervalle et tests**
- Inférence des conclusions sur la population des parcelles sur la base du modèle et à partir des données de l'échantillon des $n = 7$ parcelles

Le modèle linéaire simple (régression simple)

- On suppose que les données $y_1, \cdots, y_n$ sont
  - les réalisations de n variables aléatoires $Y_1, \cdots, Y_n$
  - liées aux quantités $x_1, \cdots, x_n$ (non aléatoires) par la relation suivante

$$\forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i$$

- où
  - x_i : valeur de x **non aléatoire** pour l'unité i
  - Y_i : réponse **aléatoire** obtenue sur l'unité i
  - ε_i **erreur résiduelle** aléatoire.
    les ε_i modélisent les erreurs de mesure, la variabilité du matériel expérimental, l'éventuelle randomisation du choix des unités,$\cdots$
  - α et β : paramètres réels inconnus (non aléatoires) : β ordonnée à l'origine et α pente de la droite de régression.

Hypothèses du modèle linéaire

$$\forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i$$

- les erreurs ε_i sont des variables aléatoires indépendantes et identiquement distribuées (iid)
- les erreurs sont centrées
- hypothèse d'homoscédasticité : les erreurs sont de même variance σ^2
- modèle **gaussien** : les erreurs sont gaussiennes $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Remarques

$$\boxed{\textbf{Modèle} \quad \forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i}$$

Dans ce modèle, Y_i se décompose en

- une **partie déterministe** $(\alpha x_i + \beta)$, **expliquée par le modèle**, et représentant l'espérance des Y_i
- une **partie aléatoire** ε_i qui reste **non expliquée par le modèle**

Objectifs

- **Contruire des estimateurs** A, B, S^2 des paramètres inconnus du modèle α, β et σ^2
- Construire des **intervalles de confiance** de α, β et de la droite de régression $\alpha x + \beta$.
- **Valider le modèle**
- **Tester le caractère significatif de la liaison linéaire**
  - Ceci revient à tester $H_0 : \alpha = 0$ contre $H_1 : \alpha \neq 0$
  - ou encore à comparer les deux modèles

$$H_0 : Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

contre

$$H_1 : Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Etudier la **qualité de l'ajustement linéaire**
- **Prédire** pour une quantité x_0 donnée, la valeur de y et préciser la confiance à accorder à cette prévision.

Construction des estimateurs

- A, B **estimateurs de α et β** obtenus par la méthode des moindres carrés

$$\begin{aligned} A &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i Y_i) - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \\ B &= \bar{Y} - A\bar{x} \end{aligned}$$

- **Estimations de α et β** : réalisations a et b des estimateurs A et B sur les données

$$\begin{aligned} a &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \\ b &= \bar{y} - a\bar{x} \end{aligned}$$

- a et b sont les coefficients de la droite des moindres carrés.

Estimateur de la variance résiduelle σ^2

- ε_i non observables
- **résidus aléatoires** $E_i = Y_i - Ax_i - B$ observables
- $\widehat{Y}_i = Ax_i + B$, la prévision (aléatoire) par le modèle de régression linéaire associée à x_i
- $E_i = Y_i - \widehat{Y}_i$.
- S^2 **estimateur de** σ^2 : variance empirique des résidus

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n} E_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - Ax_i - B)^2$$

- **estimation de** σ^2 :réalisation s^2 de S^2 sur les données

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

où e_i sont les résidus observés

Propriétés et lois des estimateurs

Loi de S^2

S^2 est un estimateur sans biais de σ^2 et on a

$$\frac{(n-2)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - Ax_i - B)^2}{\sigma^2} \sim \chi^2(n-2)$$

De plus S^2 est indépendant de A, B et $\bar{Y}$.

Propriétés et loi des estimateurs A et B

Lois de A et B

A et B sont des estimateurs sans biais et consistants de α et β. A et B suivent des lois normales d'espérance α et β, et de variance

$$
\begin{aligned}
Var(A) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
Var(B) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)
\end{aligned}
$$

On a

$$\frac{(A-\alpha)}{S_A} \sim \mathcal{T}(n-2) \text{ et } \frac{(B-\beta)}{S_B} \sim \mathcal{T}(n-2)$$

avec $S_A^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $S_B^2 = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$.

Intervalle de confiance de α et de β

- à partir de la loi de A et B estimateurs de α et β
- estimateurs par intervalle de niveau de confiance $1-\delta$ de α et β

$$[A - c_\delta S_A ; A + c_\delta S_A]$$

$$[B - c_\delta S_B ; B + c_\delta S_B]$$

où c_δ est tel que $P(|\mathcal{T}(n-2)| \leq c_\delta) = 1-\delta$
- c_δ est le quantile d'ordre $1-\frac{\delta}{2}$ de la loi de $\mathcal{T}(n-2)$
- **Intervalles de confiance** (de niveau de confiance $1-\delta$) de α et β :

$$\begin{aligned} IC_{1-\delta}(\alpha) &= [a - c_\delta s_A ; a + c_\delta s_A] \\ IC_{1-\delta}(\beta) &= [b - c_\delta s_B ; b + c_\delta s_B] \end{aligned}$$

Essais avec R

On considère la base de données `heart.disease.txt`. Il s'agit d'un extrait des données `South African Heart Disease` disponibles sur `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
On dispose des variables

- `sbp` : systolic blood pressure
- `tobacco` : cumulative tobacco (kg)
- `ldl` : low density lipoprotein cholesterol
- `adiposity` : adiposity
- `famhist` : family history of heart disease (Present, Absent)
- `typea` : type-A behavior
- `obesity` : obesity
- `alcohol` : current alcohol consumption
- `age` : age at onset
- `chd` : response, coronary heart disease

Analyse univariée

- Représentation de la base de données
  `plot(base)`
- On s'intéresse à la relation existant entre l'obésité et l'adiposité.
  `plot(adip, obesity)`
- Proposer un modèle linéaire et estimer les paramètres du modèle
  `lm(obesity~adip)`
  `summary(lm(obesity~adip))`
- Regarder la relation entre l'obésité et l'age.

Test dans le modèle de régression linéaire simple gaussien

Test du caractère significatif de la liaison linéaire

- Test de Student de la nullité de la pente de régression $H_0 : \alpha = 0$ contre $H_1 : \alpha \neq 0$
- Test de Fisher de Comparaison de modèles :

$$H_0 \ : \ modele\ M_1 : Y_i = \beta + \varepsilon_i, \quad \varepsilon_i\ iid\ \mathcal{N}(0, \sigma^2)$$

contre l'alternative

$$H_1 \ : \ modele\ M_2 : Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i\ iid\ \mathcal{N}(0, \sigma^2)$$

Test de Student de la nullité de la pente de régression

- test de $H_0 : \alpha = 0$ contre $H_1 : \alpha \neq 0$ au risque $\delta = 5\%$
- **Statistique de test** :

$$T_n = \frac{A}{S_A} \sim_{H_0} \mathcal{T}(n-2)$$

- pour un risque de 1ere espèce δ fixé acceptable (par ex $\delta = 5\%$)
  - si $p - valeur < \delta$, le test de niveau δ est significatif (liaison significative)
  - si $p - valeur > \delta$, le test de niveau δ n'est pas significatif (liaison non significative)
- **R** : dernière colonne du tableau dans

```
lm(obesity~adip)
```

Test de Fisher du caractère significatif de la linéarité

- comparer les modèles

$$\begin{array}{llll} H_0: & \text{modèle } M_1 & : & Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \ iid\ \mathcal{N}(0,\sigma^2) \\ H_1: & \text{modèle } M_2 & : & Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \ iid\ \mathcal{N}(0,\sigma^2) \end{array}$$

- **Statistique de test**

$$T_n = \frac{SCM/1}{SCR/(n-2)} \sim_{H_0} F(1, n-2)$$

  - $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ **la somme des carrés totale**
  - $SCM = \sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2$ **la somme des carrés du modèle**
  - $SCR = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 = \sum_{i=1}^{n} E_i^2$ **la somme des carrés résiduelle**

- **R** : dernière ligne dans

```
lm(obesity~adip)
```

Validation du modèle

- **adéquation** : nuages de points (x, y), $(\hat{y}, e)$
- **homoscédasticité** : nuage de points $(\hat{y}, e)$ (transformation possible des données pour stabiliser la variance)
- **indépendance** des erreurs résiduelles
  - hypothèse fondamentale du modèle linéaire
  - conditions d'obtention des données ?
  - graphe (i, e_i), test des runs
- **normalité** des erreurs résiduelles
  - hypothèse la moins importante
  - résultats asymptotiques sans normalité
  - normalité à vérifier pour petits échantillons (quelques dizaines)
  - tests de normalité (Kolmogorov-smirnov, shapiro-wilks,...) déconseillés car sensibles à la non indépendance
  - histogramme des résidus
  - QQ plot (quantiles empiriques des résidus e_i (ou normalisés) en fonction des quantiles de la gaussienne)

Validation sous R

```
plot(res$fitted, res$residuals); abline(h=0)

plot(res$fitted, obesity); abline(0,1)

plot(res$residuals); abline(h=0)

hist(res$residuals, breaks=20)

qqnorm(res$residuals); abline(0,1)
```

Prévision

- données bidimensionnelles $(x_i, y_i)_{i=1,\cdots,n}$ modélisées par un MLG

$$Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \; iid \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- **Problématique**
  - étant donnée une valeur x_0 de x pour laquelle on n'a pas observé de y_0, construire une prévision de ce y_0 non disponible
  - prévision intuitive de y_0 : $\hat{y}_0 = ax_0 + b$
  - Quel sens lui donner ? Quelle qualité ?

- Si y_0 était disponible, on lui associerait une v.a. Y_0 définie par

$$Y_0 = \alpha x_0 + \beta + \varepsilon_0, \quad \varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$$

  avec $\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_n$ indépendantes
  - $E(Y_0) = \alpha x_0 + \beta$: $\boxed{\hat{Y}_0 \text{ est un estimateur sans biais de } E(Y_0)}$

- $\hat{y}_0$ est une prévision de y_0 : **construire un intervalle de prédiction (intervalle de pari) pour Y_0**

Intervalle de prévision de Y_0

- on montre que

$$\frac{(\hat{Y}_0 - Y_0)}{\sqrt{S^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim \mathcal{T}(n-2)$$

- **Intervalle de prédiction de Y_0 de niveau $1-\delta$**

$$IP_{1-\delta}(Y_0) = \left[\hat{y}_0 - t_\delta\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)};\right.$$
$$\left.\hat{y}_0 + t_\delta\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}\right]$$

où t_δ tq $P(|\mathcal{T}(n-2)| \leq t_\delta) = 1-\delta$

- $IC_{1-\delta}(E(Y_0)) \subset IP_{1-\delta}(Y_0)$

Prevision sous R

```
new <- data.frame(adip=seq(5,50, by=5))

predict(res, new)

predict.lim<- predict(res, new, interval="confidence")

matplot(new$adip, predict.lim[,-1], col="red", lty = 1,
type = "l", ylab = "predicted y")

abline(res$coefficients[1], res$coefficients[2])
```

Exemple de régression linéaire multiple

Problématique

- On s'intéresse aux performances sportives d'enfants de 12 ans. Chaque enfant passe une dizaine d'epreuves (courses, lancers, sauts,...), et les résultats sont synthétisés dans un indice global noté Y.
- On cherche à mesurer l'incidence sur ces performances de deux variables (contrôlées) **quantitatives** : la capacité thoracique x_1 et la force musculaire x_2.
- Ces trois quantités sont repérées par rapport à une valeur de référence, notée à chaque fois 0, les valeurs positives étant associées aux bonnes performances.

Les données

pour chaque enfant i, $i = 1, \cdots, 60$ on mesure

- sa **capacité thoracique** $x_{i,1}$
- sa **force musculaire** $x_{i,2}$
- la réponse : sa **performance sportive** y_i

$x_{i,1}$	$x_{i,2}$	y_i	$x_{i,1}$	$x_{i,2}$	y_i	$x_{i,1}$	$x_{i,2}$	y_i	$x_{i,1}$	$x_{i,2}$	y_i
-8	3	42	4	5	-51	-1	-1	-87	9	4	63
9	-4	124	8	0	52	3	4	75	3	-1	-22
-7	4	-15	-2	-9	16	0	1	51	3	-6	-11
-8	-1	-132	-3	1	86	-5	3	71	-3	7	65
8	-4	10	-2	4	-55	-6	5	21	4	8	-4
0	-1	-76	-1	2	-64	-4	-6	-25	6	-8	-93
-9	-7	-24	0	8	46	1	4	4	-8	-4	14
8	-1	-25	6	1	102	-8	9	3	2	3	89
-3	-6	-120	-3	-5	-14	-8	-7	-108	0	7	4
-8	0	-69	-1	3	90	-3	-6	-46	7	5	7
-1	-7	72	7	-2	74	1	-2	41	6	-3	-66
-2	-4	-42	2	6	-60	3	2	45	4	4	104
-5	6	-30	6	-1	-56	-3	8	-76	4	-1	-46
-8	-4	-105	-9	-5	-147	-9	3	-111	1	-2	-65
-3	6	-86	-3	2	22	9	2	135	7	-4	121

Le modèle linéaire gaussien multiple

- les données y_i sont supposées être
  - les réalisations de n variables aléatoires $Y_1, \cdots, Y_n$
  - liées à la capacité thoracique $x_{i,1}$ et la force musculaire $x_{i,2}$ (non aléatoires) par la **relation linéaire**

$$\boxed{Y_i = \beta + \alpha_1 * x_{i,1} + \alpha_2 * x_{i,2} + \varepsilon_i, \quad \forall i = 1, \cdots, n}$$

- où
  - $x_{i,1}$ et $x_{i,2}$ sont la capacité thoracique et la force musculaire du **contrôlées** de l'enfant i
  - Y_i est la performance sportive **aléatoire** de l'enfant i
  - $\alpha_1, \alpha_2, \beta$ sont des **paramètres** réels inconnus,
  - les **erreurs** ε_i sont des variables aléatoires indépendantes et de même loi gaussienne $\mathcal{N}(0, \sigma^2)$.

- expliquer la variable d'intérêt Y par les deux variables explicatives x_1 et x_2 contrôlées (non aléatoires)
- en décomposant Y_i comme
  - son espérance $E(Y_i) = \beta + \alpha_1 * x_{i,1} + \alpha_2 * x_{i,2}$,
    partie fixe modélisant le type de relation envisagée entre la variable à expliquer et les variables explicatives
  - et un aléa ε_i
    partie aléatoire qui reste non expliquée par le modèle

Objectifs

- construire des estimateurs A_1, A_2, B, S^2 des paramètres $\alpha_1, \alpha_2, \beta$ et σ^2
- étudier leur loi pour bâtir des intervalles de confiance pour ces paramètres
- valider le modèle

Objectifs

- Tester le caractère significatif de la liaison (en la supposant linéaire si elle existe)

  "est-ce que la capacité thoracique et la force musculaire ont une influence significative sur la performance sportive ?"
- on testera
  - $$H_0 : \alpha_1 = 0 \text{ et } \alpha_2 = 0 \quad \text{contre } H_1 : \alpha_1 \neq 0 \text{ ou } \alpha_2 \neq 0$$
  - ce qui revient à comparer les deux modèles

    $$Y_i \; = \; \beta \; + \; \varepsilon_i$$

    et

    $$Y_i \; = \; \beta \; + \; \alpha_1 * x_{i,1} \; + \; \alpha_2 * x_{i,2} \; + \; \varepsilon_i$$

Objectifs

- Tester le caractère significatif de l'influence d'une variable (en la supposant linéaire si elle existe)
  "est-ce que la capacité thoracique a une influence significative en plus de la force musculaire poids sur la performance sportive ?"
- On testera
  - $$H_0 : \alpha_1 = 0 \;\; \text{contre} \;\; H_1 : \alpha_1 \neq 0$$
  - ce qui revient à comparer les deux modèles
    $$Y_i \;=\; \beta \;+\; \alpha_2 * x_{i,2} \;+\; \varepsilon_i$$
    et
    $$Y_i = \beta + \alpha_1 * x_{i,1} + \alpha_2 * x_{i,2} + \varepsilon_i$$
- Prédire et apprécier la qualité de la prédiction fournie en calculant un intervalle de prévision

Le modèle de régression linéaire multiple

- $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})_{1 \leq i \leq n}$ mesures de p variables explicatives
- $y_1, \cdots, y_n$ réalisations de $Y_1, \cdots, Y_n$ liées aux $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ par la relation

$$\forall i = 1, \cdots, n \quad Y_i = \beta + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \ldots + \alpha_p x_{i,p} + \varepsilon_i$$

- $\alpha_1, \ldots, \alpha_p$ et β : paramètres d'espérance, appelés coefficients de régression
- α_j représente l'accroissement de Y_i correspondant à l'accroissement d'une unité de x_j quand les autres variables explicatives sont fixées
- ε_i erreurs résiduelles aléatoires (bruits que le modèle n'explique pas) supposées iid de loi $\mathcal{N}(0, \sigma^2)$
- paramètre de variance : σ^2

Ecriture matricielle du MLGM

$$\left\{\begin{array}{lllllllllllllll} Y_1 & = & \beta & + & \alpha_1\, x_{1,1} & + & \alpha_2\, x_{1,2} & + & \ldots & + & \alpha_p\, x_{1,p} & + & \varepsilon_1 \\ Y_2 & = & \beta & + & \alpha_1\, x_{2,1} & + & \alpha_2\, x_{2,2} & + & \ldots & + & \alpha_p\, x_{2,p} & + & \varepsilon_2 \\ \vdots & & & & & & & & & & & & \\ Y_n & = & \beta & + & \alpha_1\, x_{n,1} & + & \alpha_2\, x_{n,2} & + & \ldots & + & \alpha_p\, x_{n,p} & + & \varepsilon_n \end{array}\right.$$

peut se réécrire sous la forme

$$Y \;=\; X\,\theta \;+\; \varepsilon$$

avec

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\; X = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ 1 & x_{2,1} & \ldots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{pmatrix},\; \theta = \begin{pmatrix} \beta \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix},\; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Estimateurs des paramètres d'espérance

- Estimateur des moindres carrés de θ

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\,\theta\|^2$$

c'est la valeur de θ qui réalise le minimum de $\|Y - X\,\theta\|^2$

- $\hat{\theta} = (B, A_1, \ldots, A_p)^t$ est solution du système

$$(X^t\,X)\,\hat{\theta} = X^t\,Y$$

- si $(X^t\,X)$ est inversible,

$$\hat{\theta} = (X^t\,X)^{-1}\,X^t\,Y$$

- estimations des paramètres $\beta, \alpha_1, ...\alpha_p$ sont les réalisations $b, a_1, \cdots, a_p$ des estimateurs $B, A_1, \ldots, A_p$ sur l'échantillon

Estimateur de la variance résiduelle σ^2

- S^2 **estimateur de σ^2** :

$$\begin{aligned} S^2 &= \frac{1}{n-p-1} \sum_{i=1}^{n} E_i^2 \\ &= \sum_{i=1}^{n} \left(Y_i - (B + \sum_{j=1}^{p} A_j x_{i,j}) \right)^2 = \frac{\|Y - X\hat{\theta}\|^2}{n-p-1} \end{aligned}$$

- **estimation de σ^2** :réalisation s^2 de S^2 sur les données

$$s^2 = \sum_{i=1}^{n} (y_i - (b + a_1 x_{i,1} + \cdots + a_p x_{i,p}))^2$$

Intervalle de confiance de θ_k

- à partir de la loi de $\hat{\theta}_k$ estimateur de θ_k
- estimateurs par intervalle de niveau de confiance $1-\delta$ de chaque coefficient de régression θ_k

$$\left[\hat{\theta}_k - s_\delta \sqrt{S^2 c_{kk}} ; \hat{\theta}_k + s_\delta \sqrt{S^2 c_{kk}}\right]$$

- s_δ quantile d'ordre $1-\frac{\delta}{2}$ de la $\mathcal{T}(n-p-1)$

Modèle multiple sous R

Reprendre les données de la base `heart.disease`

- Modèle à deux variables explicatives
  ```
  resM<- lm(obesity~adip+age)
  summary(resM)
  ```
- Modèle à plus de deux variables explicatives
  ```
  resM<- lm(obesity~sbp+tobacco+ldl+adip+alchohol+age)
  ```

1. Test de la contribution globale des variables explicatives

- **Premier test à effectuer**

  H_0 : aucune des p variables explicatives n'a d'influence sur Y contre

  H_1 : au moins une des variables explicatives contribue à expliquer Y
- c'est à dire Test du modèle constant

$$H_0 \ : \ modele \ M_1 : Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \ iid \ \mathcal{N}(0, \sigma^2)$$

contre le modèle complet

$$H_1 \ : \ modele \ M_{p+1} : Y_i = \beta + \sum_{j=1}^{p} \alpha_j x_{i,j} + \varepsilon_i, \quad \varepsilon_i \ iid \mathcal{N}(0, \sigma^2)$$

soit encore

- Test de H_0 : $\forall j = 1, \cdots, p, \ \alpha_j = 0$ contre $H_1 : \exists j \ \alpha_j \neq 0$
- Test de Fisher de loi $F(p, n-p-1)$

2. Test du modèle M_{q+1} contre le modèle M_{p+1}

- tester si un ensemble de q variables explicatives ne suffit pas à expliquer Y
- c'est à dire Test de

$$H_0 \ : \ modele\ M_{q+1} \ : Y_i = \beta + \sum_{j=1}^{q} \alpha_j x_{i,j} + \varepsilon_i, \quad \varepsilon_i\ iid \mathcal{N}(0, \sigma^2)$$

contre l'alternative

$$H_1 \ : \ modele\ M_{p+1} : Y_i = \beta + \sum_{j=1}^{p} \alpha_j x_{i,j} + \varepsilon_i, \quad \varepsilon_i\ iid \mathcal{N}(0, \sigma^2)$$

soit encore

- Test de $H_0 : \ \forall j = q+1, \cdots, p, \ \alpha_j = 0$ contre $H_1 : \exists j = q+1, \cdots, p\ \alpha_j \neq 0$ au risque δ
- Test de Fisher de loi $F(p-q\,;\, n-p-1)$

Sous R

- Test global : `summary( lm(obesity~adip+age))`
- Test de deux modèles emboités

```
resM1<- lm(obesity~adip+age)
resMC<- lm(obesity~sbp+tobacco+ldl+adip+alchohol+age)
anova(resM1,resMC)
```

Sélection de variables et sélection de modèles

- si le test global $H_0 : M_1$ contre $H_1 : M_{p+1}$ est significatif, au moins une des variables contribue à expliquer Y
- quelles variables contribuent réellement à expliquer Y parmi les $x_1, \cdots, x_p$?
- **première idée fausse** :
  - tester la nullité de chaque coefficient de régression avec le test de Student $\forall k = 1, \cdots, p$, $H_0 : \alpha_k = 0$ contre $H_1 : \alpha_k \neq 0$
  - éliminer toutes les variables x_k tq le test de Student associé n'est pas significatif
  - démarche fausse : chaque test est effectué alors que les autres variables explicatives sont fixées, on ne prend pas en compte les possibles effets conjoints

2ème idée : Sélectionner les variables pertinentes

- **méthode de recherche exhaustive**
  - nécessité de comparer 2^p modèles
  - si p pas trop élevé, on peut comparer tous les modèles possibles et choisir "le meilleur"' modèle à partir d'un critère statistique de sélection de modèles
  - attention : le test de Fisher ne permet de comparer que des modèles emboités
- **Méthode de sélection de variables pas à pas**
  - extraire un groupe de variables suffisamment explicatif
  - conserver un **modèle explicatif** : relativement simple et facile à interpréter
  - introduire ou supprimer une variable l'une après l'autre
  - pas de garantie de résultat optimal
  - ne met pas à l'abri d'enlever des variables réellement significatives

Méthodes de sélection de variables

Méthode ascendante (forward)

- Modèle sans covariable
- Insertion de la variable qui explique le plus Y et qui est significative
- Insertion de la 2eme variable qui explique le plus Y et qui est significative, etc

Méthode descendante (backward)

- Modèle complet
- Enlever la variable qui explique le moins Y et qui est NS
- Enlever la 2eme variable qui explique le moins Y et est NS, etc

Méthode pas à pas (stepwise)

- ascendante avec remise en cause à chaque étape des variables déjà introduites
- permet d'éliminer les variables qui ne sont plus informatives compte tenu de celle qui vient d'etre sélectionnée.

Sous R

- Modèle complet
```
resMC<- lm(obesity~sbp+tobacco+ldl+adip+alchohol+age)
```
- Selection descendante
```
step(resMC,direction="backward")
```
- Selection pas à pas
```
step(resMC, direction = "both")
```
- Modèle final
```
resF<-lm(obesity~adip+age)
summary(resF)
```

Validation du modèle

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\vec{0}_n\,,\, \sigma^2\, I_n)$$

Vérifier les hypothèses du modèle :

- adéquation : $\forall i, E(\varepsilon_i) = 0$;
- homoscédasticité $V(\varepsilon_i) = \sigma^2$, $\forall i$
- indépendance des erreurs résiduelles
- normalité des erreurs résiduelles

à partir des résidus $e_1, \ldots, e_n$

Validation du modèle

- **adéquation** : nuage de points $(\hat{y}, e)$ dans lequel les résidus ne doivent présenter aucune propriété intéressante
- **homoscédasticité** : nuage de points $(\hat{y}, e)$ (transformation possible des données pour stabiliser la variance)
- **indépendance** des erreurs résiduelles
  - hypothèse fondamentale du modèle linéaire
  - graphe (i, e_i) l'ordre des résidus ne doit pas avoir de sens
- **normalité** des erreurs résiduelles
  - hypothèse la moins importante
  - normalité à vérifier pour petits échantillons (quelques dizaines)
  - tests de normalité (Kolmogorov-smirnov, shapiro-wilks,...) déconseillés car sensibles à la non indépendance
  - histogramme des résidus, QQ plot (quantiles empiriques des résidus e_i (ou normalisés) en fonction des quantiles de la gaussienne)
- **graphiques partiels** : tracé des p nuage de points (x_k, e) pour chaque variable explicative pour traquer les dépendances entre résidus et variables explicatives et détecter les points atypiques et/ou influents

Validation sous R

```
plot(resF$fitted, resF$residuals); abline(h=0)

plot(resF$fitted, obesity); abline(0,1)

plot(resF$residuals); abline(h=0)

hist(resF$residuals, breaks=20)

qqnorm(resF$residuals); abline(0,1)
```