Modélisation statistiques

aurore.lavigne@univ-lille.fr

Organisation du cours

- COURS : Mme Lavigne le vendredi de 8h30 à 10h00 => 10 séances
- TD => 10,5 séances
  - Mme Lavigne le lundi de 15h15 à 17h15
  - Mme GUin le vendredi de 14h45 à 16h45
- Evaluations : $NF = 0.75DS + 0.25CC$, le rattrapage remplace toutes les notes.

  DS deux devoirs surveillés

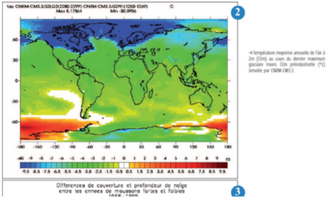
  CC note de TD, DM, participation
- Moodle : code d'inscription rapide : 6j46pw
- Références parmi d'autres :
  - SAPORTA, Gilbert. Probabilités, analyse des données et statistique. Editions Technip, 2006.
  - DAUDIN, Jean-Jacques. Le modèle linéaire et ses extensions-Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (Niveau C). 2015.

Plan du cours

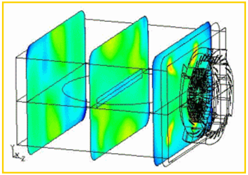
Qu'est-ce que la modélisation ?

Les modèles dans la vie quotidienne

- Modèle METEO ou CLIMAT
- Dans l' industrie : ex, répartition de la chaleur dans un four
- Pharmaceutique : évolution de la concentration d'une certaine molécule dans le sang après ise d'un médicament.
- En économie
- En sciences cognitives



Source : Cnrm



Source : SOLSI-CAD

Qu'est ce qu'un modèle ?

- représente de manière formelle (à l'aide d'équations mathématiques) la réalité d'un phénomène.

Qu'est ce qu'un modèle ?

- représente de manière formelle (à l'aide d'équations mathématiques) la réalité d'un phénomène.
- simplification de la réalité : ils sont tous un peu faux.

Qu'est ce qu'un modèle ?

- représente de manière formelle (à l'aide d'équations mathématiques) la réalité d'un phénomène.
- simplification de la réalité : ils sont tous un peu faux.
- Intérêt : calcul de valeurs d'intérêt
  - Quel sera la météo de demain ? => prédiction
  - Quel sera le climat dans 100 ans ?
  - La température du four est-elle homogène ?
  - Combien doit produire une entreprise pour maximiser son profit ?

Qu'est ce qu'un modèle ?

- représente de manière formelle (à l'aide d'équations mathématiques) la réalité d'un phénomène.
- simplification de la réalité : ils sont tous un peu faux.
- Intérêt : calcul de valeurs d'intérêt
  - Quel sera la météo de demain ? => prédiction
  - Quel sera le climat dans 100 ans ?
  - La température du four est-elle homogène ?
  - Combien doit produire une entreprise pour maximiser son profit ?
- Tous ces modèles ont une unique solution : ils sont déterministes. Ils sont généralement écrits à l'aide d'équations différentielles.

Modèlisation statistiques

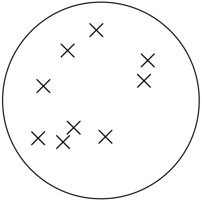
Contexte

- Lorsqu'on répète l'expérience, le résultat peut varier
- La modélisation statistique s'appuie sur les résultats d'une expérience :
  - résultats d'une expérimentation (ex temps de réaction pour réagir à un stimuli)
  - réponses de personnes sondées : aléa provient du choix de l'échantillon
  - mesure de la qualité d'un échantillon de pièces dans une production
- On apellera ces "mesures" des observations ou des données.

Ojectif

- Décrire le phénomène - processus à l'origine des données.
- Celui-ci est aléatoire : on utilisera des lois de probabilités.
- Deux rôles :
  - Descriptif : comprendre comment sont générés les observations, quel lien il peut y avoir entre les variables.
  - Prédictif : pouvoir faire une prédiction pour une nouvelle date, lieu, individu...

Schéma de l'inférence statistique



Population

Schéma de l'inférence statistique

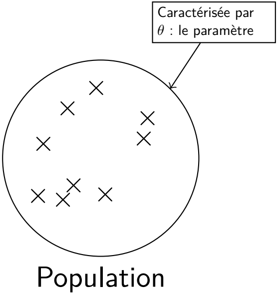


Schéma de l'inférence statistique

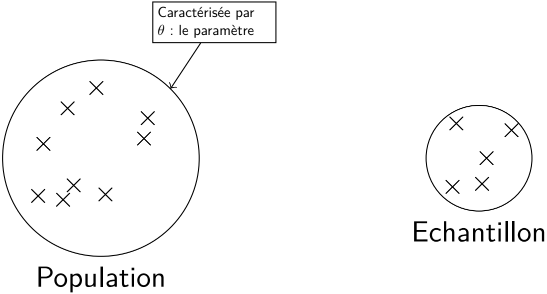


Schéma de l'inférence statistique

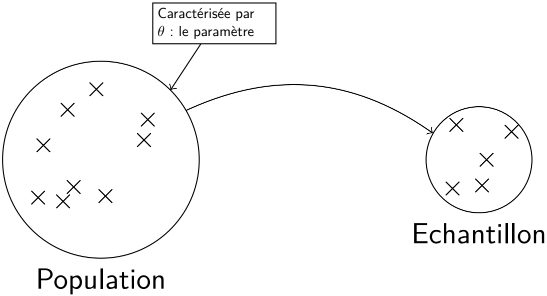


Schéma de l'inférence statistique

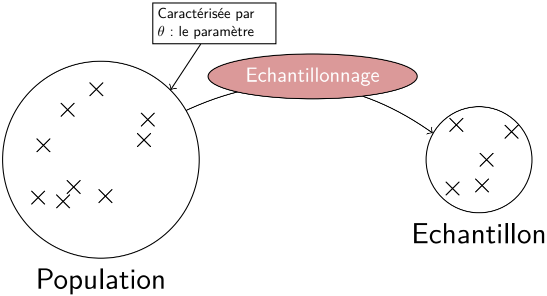


Schéma de l'inférence statistique

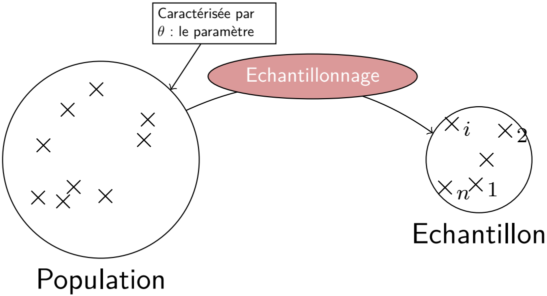


Schéma de l'inférence statistique

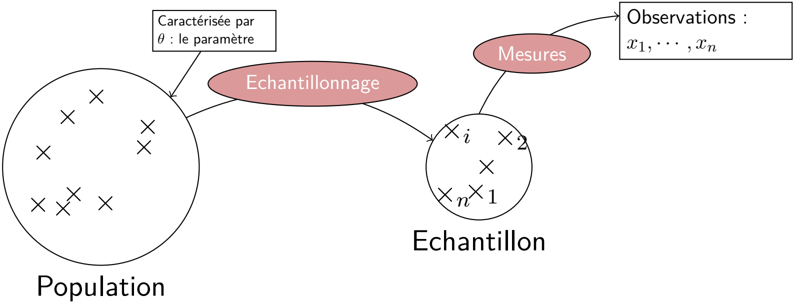


Schéma de l'inférence statistique

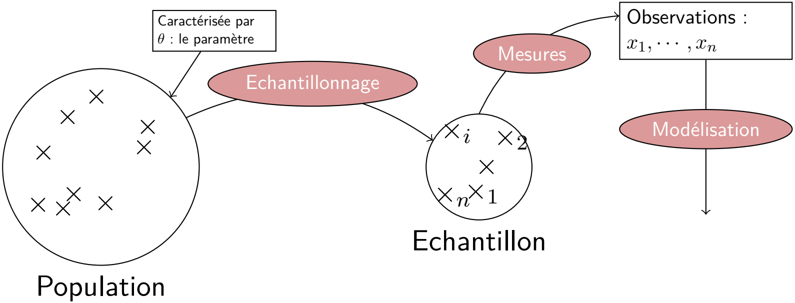


Schéma de l'inférence statistique

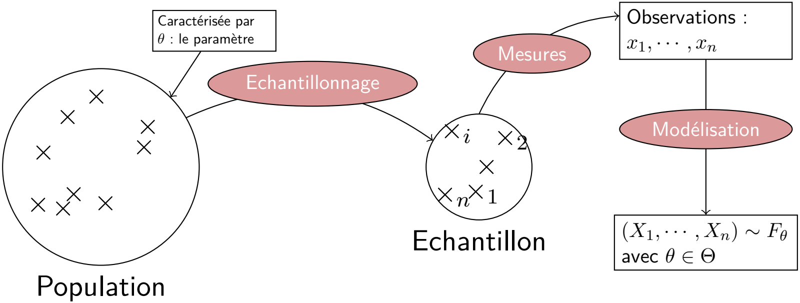


Schéma de l'inférence statistique

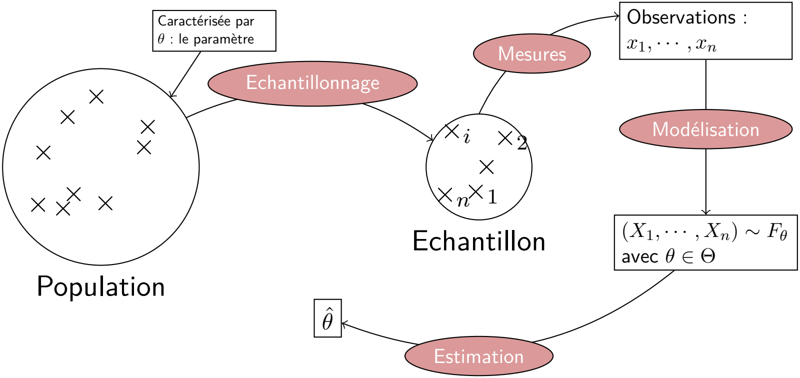
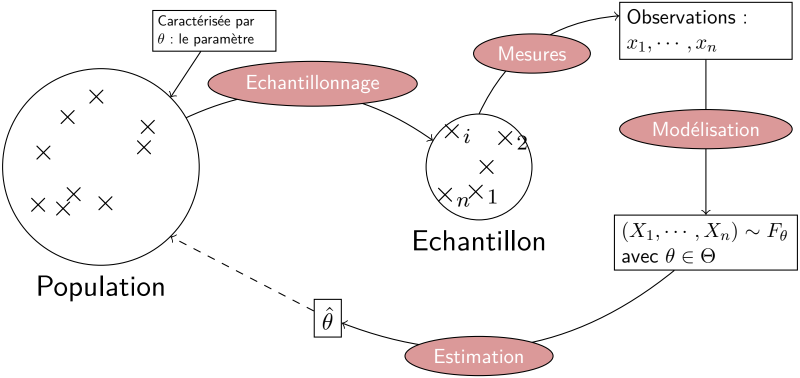


Schéma de l'inférence statistique



La modélisation statistique c'est

1. Supposer que les observations $x_1, x_2, \cdots, x_n$ sont la réalisation de variables aléatoires $X_1, X_2, \cdots, X_n$.

2. Donner une famille de loi possible $\mathcal{F}$ pour le vecteur aléatoire $(X_1, X_2, \cdots, X_n)$
   2.1 Si on peut écrire $\mathcal{F}$ sous la forme suivante $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$ avec θ de dimension fini, on dira que le modèle est paramétrique. Par exemple $\mathcal{F} = \{\mathcal{P}(\lambda), \lambda \in \mathbb{R}^+\}$.
   2.2 On considèrera souvent que les variables $X_1, \cdots, X_n$ sont *i.i.d.*.
      i. indépendantes, les variables X_i sont indépendantes,
      i.d. identiquement distribuées, toutes les variables X_i suivent la même loi.

   Dans ce cas, il suffit de donner la loi d'un X_i pour spécifier le modèle complètement.

Exemple : échantillonnage aléatoire simple

- Echantillonnage On considère une population de très grande dans taille dans laquelle on tire n individus sans remises. L'échantillonnage aléatoire simple assure à chaque individu d'avoir la même probabilité d'être tirés.

Exemple : échantillonnage aléatoire simple

- Echantillonnage On considère une population de très grande dans taille dans laquelle on tire n individus sans remises. L'échantillonnage aléatoire simple assure à chaque individu d'avoir la même probabilité d'être tirés.
- Mesure Pour chaque individu de l'échantillon, on note son sexe. On écrira $x_i = 0$, si le i^{eme} individu de l'échantillon est un homme, $x_i = 1$ si c'est une femme.

Exemple : échantillonnage aléatoire simple

- Echantillonnage On considère une population de très grande dans taille dans laquelle on tire n individus sans remises. L'échantillonnage aléatoire simple assure à chaque individu d'avoir la même probabilité d'être tirés.
- Mesure Pour chaque individu de l'échantillon, on note son sexe. On écrira $x_i = 0$, si le i^{eme} individu de l'échantillon est un homme, $x_i = 1$ si c'est une femme.
- Modélisation On supposera que les X_i sont indépendants, cela est possible car on a réalisé un échantillonnage aléatoire simple, et que la taille de la population est très grande devant celle de l'échantillon. On posera alors, X_i représente le sexe du i^{eme} individu de l'échantillon,

  $$X_i \sim \mathcal{B}(p) \text{i.i.d.}$$

  p est le paramètre à estimer, c'est la proportion de femmes dans la population.

Exemple : échantillonnage aléatoire simple

- Echantillonnage On considère une population de très grande dans taille dans laquelle on tire n individus sans remises. L'échantillonnage aléatoire simple assure à chaque individu d'avoir la même probabilité d'être tirés.
- Mesure Pour chaque individu de l'échantillon, on note son sexe. On écrira $x_i = 0$, si le i^{eme} individu de l'échantillon est un homme, $x_i = 1$ si c'est une femme.
- Modélisation On supposera que les X_i sont indépendants, cela est possible car on a réalisé un échantillonnage aléatoire simple, et que la taille de la population est très grande devant celle de l'échantillon. On posera alors, X_i représente le sexe du i^{eme} individu de l'échantillon,

$$X_i \sim \mathcal{B}(p)\text{i.i.d.}$$

  p est le paramètre à estimer, c'est la proportion de femmes dans la population.
- Estimation

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Les méthodes d'échantillonnage 1/2

En population finie, on veut rendre compte des paramètres de la population à travers un échantillon. Il existe différentes méthodes d'échantillonnage. Ce n'est qu'en controlant votre métode d'échantillonnage que vous pourrez dire des choses sur la population.

- L'échantillonnage aléatoire simple Tous les individus de la population ont la même probabilité d'être tirés. Attention, à utiliser une méthode qui vous assure vraiment cela, sinon vous pourrez introduire un bien (cf TD).
- L'échantillonnage aléatoire stratifié Il est utile si il existe des groupes d'individus plus ou moins homogènes dans votre population, et que pour chaque individu de la population vous connaissee son groupe. Dans ce cas, on fixe un nombre d'individus à tirer dans chaque groupe et on oprère à un échantillonnage aléatoire simple dans chacun des groupes. Cette méthode peut vous permettre de diminuer la variance de l'estimateur.

Les méthodes d'échantillonnage 2/2

- L'échantillonnage en grappes Là aussi il existe des groupes dans la population. Au lieu de tirer les individus vous procédez à un échantillonnage aléatoire simple sur les groupes, et vous prenez dans les l'échantillon tous les individus des groupes sélectionnés. Cela peut vous permettre de baisse le coût de l'étude.

Pour en savoir plus : DROESBEKE, Jean-Jacques. *Les sondages*. FeniXX, 1987.

Plan du cours

Plan du cours

Contexte

On dispose d'un couple de variables aléatoires (X, Y).

Théorème de la variance totale

$$\underbrace{V(Y)} \quad = \quad \underbrace{E[V(Y)|X]} \quad + \quad \underbrace{V(E[Y|X])}$$

Et donc

$$V(Y) \leq E[V(Y)|X]$$

Le fait de connaître X permet de diminuer l'incertitude sur Y. Ainsi X pourrait servir à prédire Y. On pourrait prédire Y par une fonction de X : $\hat{Y} = f(X)$.

Contexte

On dispose d'un couple de variables aléatoires (X, Y).

Théorème de la variance totale

$$\underbrace{V(Y)}_{\text{Variance totale}} = \underbrace{E[V(Y)|X]}_{\text{Variance résiduelle}} + \underbrace{V(E[Y|X])}_{\text{Variance expliquée}}$$

Et donc

$$V(Y) \leq E[V(Y)|X]$$

Le fait de connaître X permet de diminuer l'incertitude sur Y. Ainsi X pourrait servir à prédire Y. On pourrait prédire Y par une fonction de X : $\hat{Y} = f(X)$.

Contexte

Résultat important

$$\text{argmax} E((Y - f(X))^2) = E(Y|X)$$

La meilleure façon d'utiliser X pour prédire Y est de prendre $f(X) = E(Y|X)$. On va lors poser

$$Y = E(Y|X) + \epsilon$$

avec ϵ un terme d'erreur aléatoire. On peut montrer que :

- $E(\epsilon) = 0$
- $cov(\epsilon, X) = cov(\epsilon, E[Y|X]) = 0$
- $V(\epsilon) = (1 - \frac{V(E(Y|X))}{V(Y)})V(Y)$

Cadre de la régression linéaire

Dans le cadre de la régression linéaire on va poser

$$E(Y|X) = \alpha + \beta X$$

c'est à dire

$$Y = \alpha + \beta X + \epsilon$$

Il existe une unique droite satisfaisant $E(Y|X) = \alpha + \beta X$. Cette droite à pour équation :

$$Y = E(Y) + \rho \frac{\sigma_Y}{\sigma_X}(X - E(X)) + \epsilon$$

Application à la statistique

On dispose d'un échantillon de taille n, sur lequel on mesure deux variables que l'on nommera y_i et x_i.

On supposera que le couple (x_i, y_i) est la réalisation d'un couple aléatoire (X_i, Y_i). On supposera également que tous les couples (X_i, Y_i) sont indépendants et identiquement distribués.

On se demande si la variable X_i à une influence sur la variable Y_i.

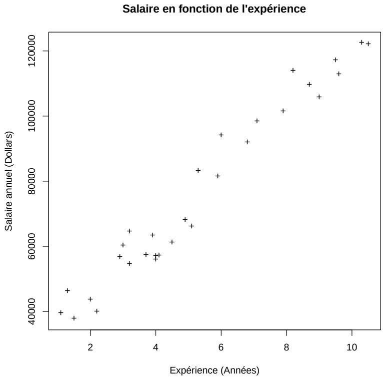
On appelera les deux variables ainsi

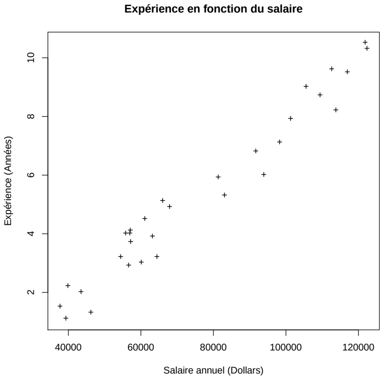
Y_i	X_i
variable endogène	variable exogène
variable dépendante	variable indépendante
variable à expliquer	variable explicative

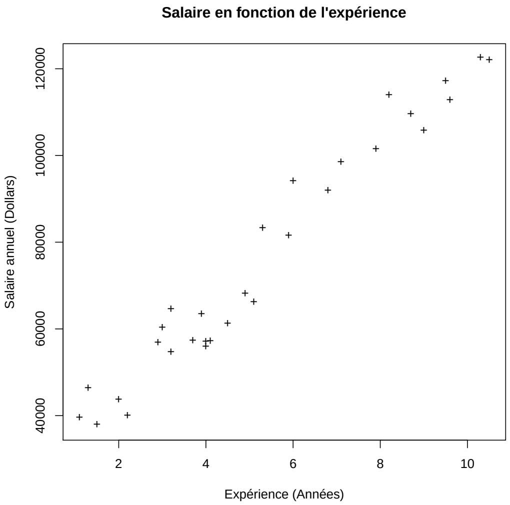
L'exemple des salaires

Pour $n = 30$ salariés d'un entreprise, on dispose de leur salaire annuel et de leur années d'expérience.

YearsExperience	Salary
1.10	39343.00
1.30	46205.00
1.50	37731.00
2.00	43525.00
2.20	39891.00
2.90	56642.00
3.00	60150.00
3.20	54445.00
3.20	64445.00
3.70	57189.00
3.90	63218.00
4.00	55794.00
4.00	56957.00


Salaire en fonction de l'expérience
Salaire annuel (Dollars)
40000
60000
80000
100000
120000
2
4
6
8
10
Expérience (Années)


Expérience en fonction du salaire
Expérience (Années)
2
4
6
8
10
40000
60000
80000
100000
120000
Salaire annuel (Dollars)



Comment la variable "Expérience" *explique* la variable "Salaire" ?

Le modèle linéaire

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$

- Y_i Variable dépendante pour le i^{eme} individu de l'échantillon
  => ALÉTAOIRE
- x_i Variable indépendante pour le i^{eme} inidividu de l'échantillon.
  => NON ALÉATOIRE
- E_i terme d'erreur pour le i^{eme} individu de l'échantillon.
  => ALÉTAOIRE
  - Les erreurs sont supposées **indépendantes**.
  - Les erreurs sont supposées **homoscédastiques**, la variance est identique quelque soit l'individu : σ^2.

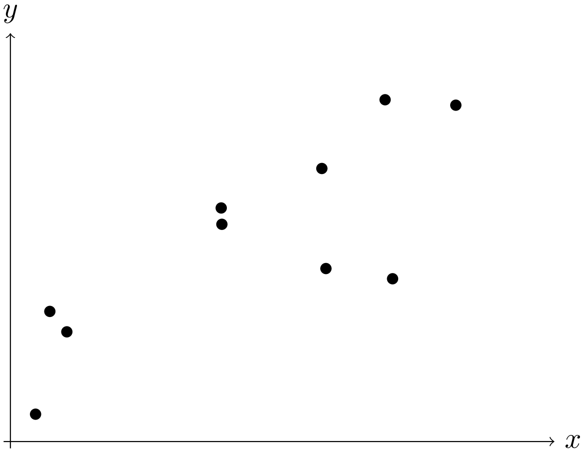
$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$

- α, β sont les paramètres de régression à estimer.
  => NON ALÉATOIRE
  - α est l'ordonnée à l'origine, il a la même unité que Y_i
  - β est la pente, il s'exprime en unité de Y_i par unité de x_i. β représente l'augmentation ou la diminution de Y par unité d'augmentation de X.
- σ^2 est le paramètre de variance à estimer.
  => NON ALÉATOIRE

REMARQUE :

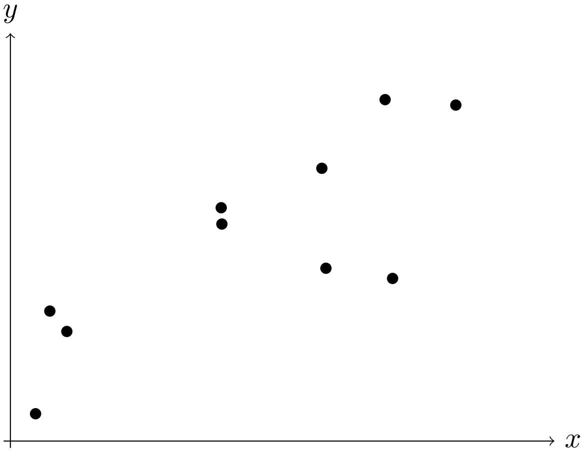
- La variable indépendante est considérée comme fixe (non aléatoire). C'est en pratique le cas, dans les expériementations, lorsqu'on fixe les conditions initiales. Les résultats des moindres carrés sont les mêmes que l'on considère X comme aléatoire ou non. C'est ce qui distingue le modèle de **regréssion** du modèle **linéaire**.

Estimation de α, β et σ par les moindres carrés.



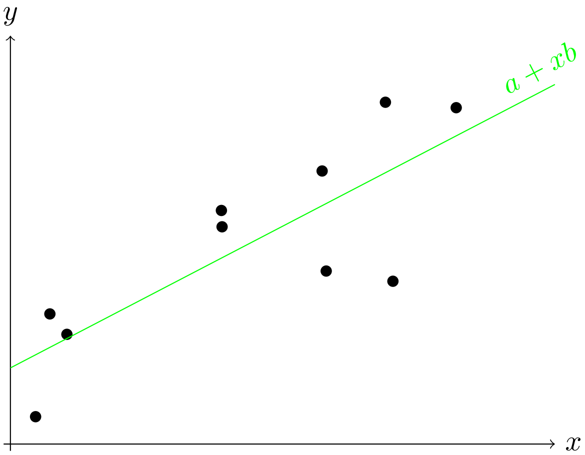
a et b définissent la droite d'ajustement.

Estimation de α, β et σ par les moindres carrés.



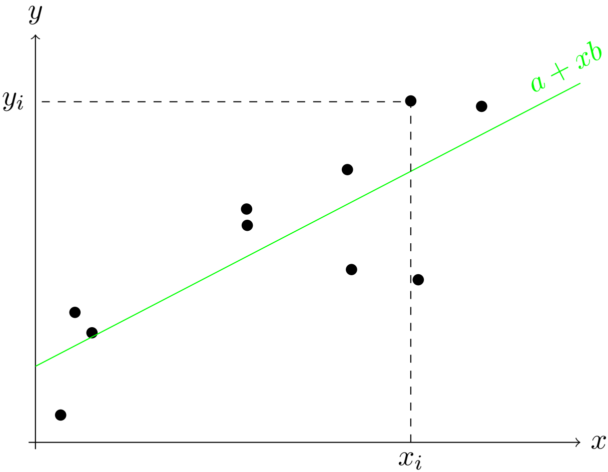
a et b définissent la droite d'ajustement.

Estimation de α, β et σ par les moindres carrés.



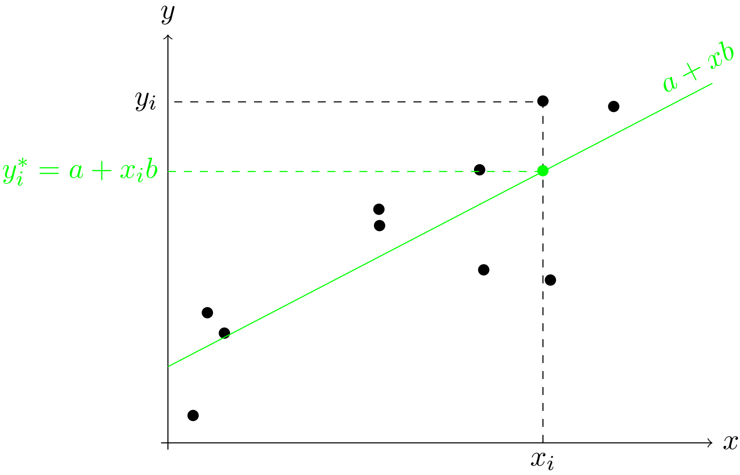
a et b définissent la droite d'ajustement.

Estimation de α, β et σ par les moindres carrés.



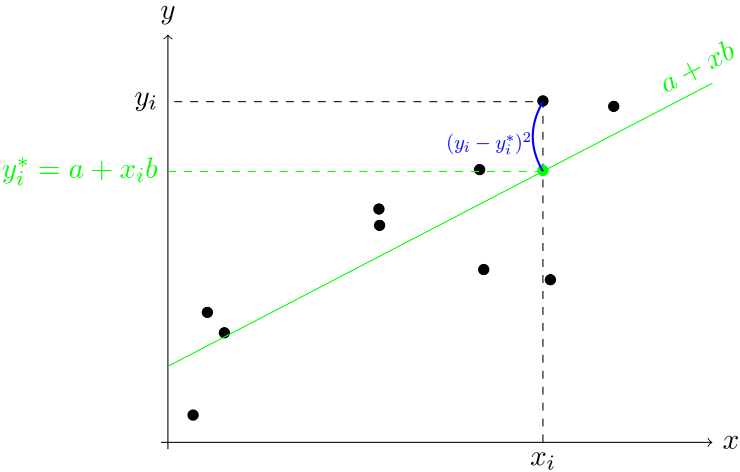
La méthode des moindres carrés ordinaires (MCO), consiste à choisir a et b qui minimisent $\sum_{i=1}^{n} (y_i - (\alpha + x_i\beta))^2$

Estimation de α, β et σ par les moindres carrés.



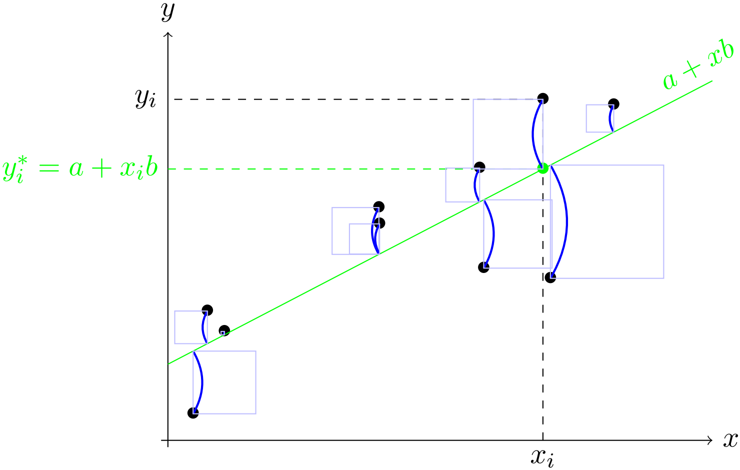
La méthode des moindres carrés ordinaires (MCO), consiste à choisir a et b qui minimisent $\sum_{i=1}^{n} (y_i - (\alpha + x_i\beta))^2$

Estimation de α, β et σ par les moindres carrés.



La méthode des moindres carrés ordinaires (MCO), consiste à choisir a et b qui minimisent $\sum_{i=1}^{n} (y_i - (\alpha + x_i\beta))^2$

Estimation de α, β et σ par les moindres carrés.



La méthode des moindres carrés ordinaires (MCO), consiste à choisir a et b qui minimisent $\sum_{i=1}^{n} (y_i - (\alpha + x_i\beta))^2$

Solutions : les équations du premier ordre

On pose la fonction de coût suivante

$$F(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha + \beta x_i)^2$$

Selon la méthode des moindres carrés ordinaires on estime α et β par

$$(a, b) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\text{argmin}}\, F(\alpha, \beta)$$

(a, b) est la solution du système **d'équations normales**

$$\begin{cases} \frac{\partial F}{\partial \alpha}(\alpha, \beta) = 0 \\ \frac{\partial F}{\partial \beta}(\alpha, \beta) = 0 \end{cases} \Leftrightarrow \begin{cases} \overline{y} = \alpha + \beta\overline{x} \quad (1) \\ \sum_{i=1}^{n} x_i(y_i - (\alpha + \beta x_i)) = 0 \quad (2) \end{cases}$$

avec $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Solutions

$$b = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$a = \overline{y} - \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\overline{x}$$

REMARQUES :

- b est la version empirique de $\frac{Cov(X,Y)}{V(X)}$ qui est la pente de la droite d'équation $E(Y|X) = \alpha + \beta X$.
- a et b dépendent des y_i qui sont des réalisations des variables aléatoires Y_i, a et b sont donc les réalisations de variables aléatoires.
- On pose $y_i^* = a + bx_i = \overline{y} + b(x_i - \overline{x})$, c'est l'estimation de $E(Y_i|X = x_i)$
- La droite estimée passe par le point de gravité du nuage de points, de coordonnée $(\overline{x}, \overline{y})$.

Propriétés

a, b et y_i^* sont des estimations sans biais de, respectivement, α, β et $E[Y|X = x_i] = \alpha + \beta x_i$.

- b est une réalisation de $B = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$ et $E(B|X_1 = x_1, \cdots, X_n = x_n) = \beta$
- a est une réalisation de $A = \overline{Y} - B\overline{X}$ et $E(A|X_1 = x_1, \cdots, X_n = x_n) = \alpha$
- y_i^* est une réalisation de $Y_i^* = A + Bx_i$ et $E(Y_i^*|X_1 = x_1, \cdots, X_n = x_n) = E(Y_i|X = x_i) = \alpha + \beta x_i$
- B n'est pas corrélé à $\overline{Y}$, à x_i fixé.

Variance de estimateurs

Propriété de Gauss-Markov

Parmi les estimateurs linéaires des Y_i, A et B sont de variance minimales.

On dit que ce sont des estimateurs BLUE *Best Linear Unbiaised Estimates*

Preuve : plus tard.

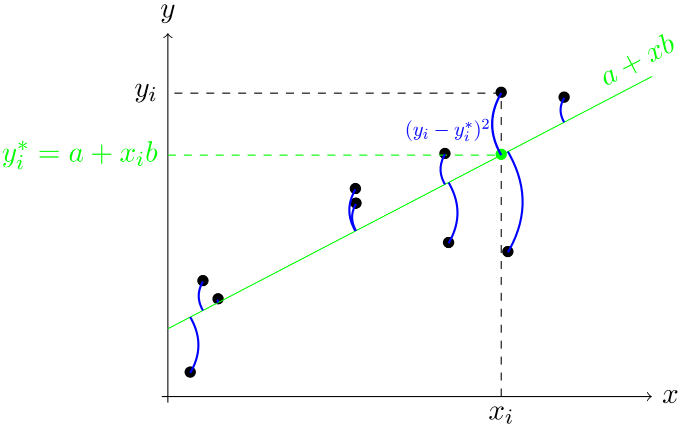
Variance des estimateurs

$$V(B) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad V(A) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)$$

$$V(Y_i^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)$$

Et σ^2 ?

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$



Et σ^2 ?

- Dans le modèle, σ^2 est la variance des erreurs,
- puisque Y_i^* est un estimateur de $\alpha + \beta x_i$, on va utiliser la variance des résidus $E_i^* = Y_i - Y_i^*$ pour estimer σ^2.

On montrera (plus tard) que

$$S^2 = \frac{\sum_{i=1}^n (Y_i - Y_i^*)^2}{n-2}$$

est un estimateur sans biais de σ^2.

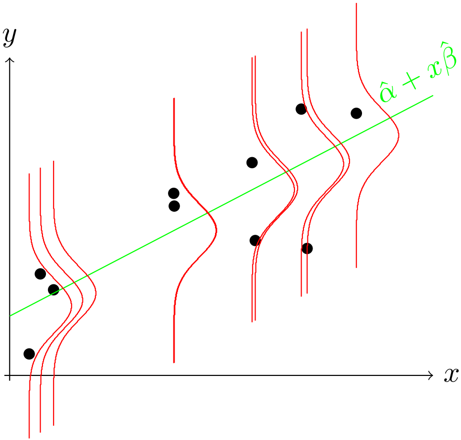
Propriétés des résidus

- Les résidus $e_i = y_i - y_i^*$ sont de moyenne nulle.
- Les résidus ne sont donc pas des réalisations de variables aléatoires indépendantes (comme sur le modèle).
- On note $s_n^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ la variance empirique des résidus.
- On a $s_n^2 = (1 - r^2)s_y^2$ avec r le coefficient de corrélation empirique, défini par

$$r = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (x_i - \overline{x})^2}}$$

Le modèle linéaire gaussien

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \sim \mathcal{N}(0, \sigma^2) \quad i.i.d.$$



A partir de maintenant, on ne considère plus que ce modèle.

Le modèle linéaire

On a les résultats suivants :

- $Y_i|X_i = x_i \sim \mathcal{N}(\alpha + \beta x, \sigma^2)$ i.i.d.
- B, A et Y_i^* suivent des lois normales, comme combinaisons linéaires de variables aléatoires normales indépendantes.

$$B \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2})$$

$$A \sim \mathcal{N}(\alpha, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}))$$

$$Y_i^* \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}))$$

- Selon le théorème de Cochran

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

- S^2 est indépendant de $\overline{Y}$, B et A.

Application au cas des salaires avec R

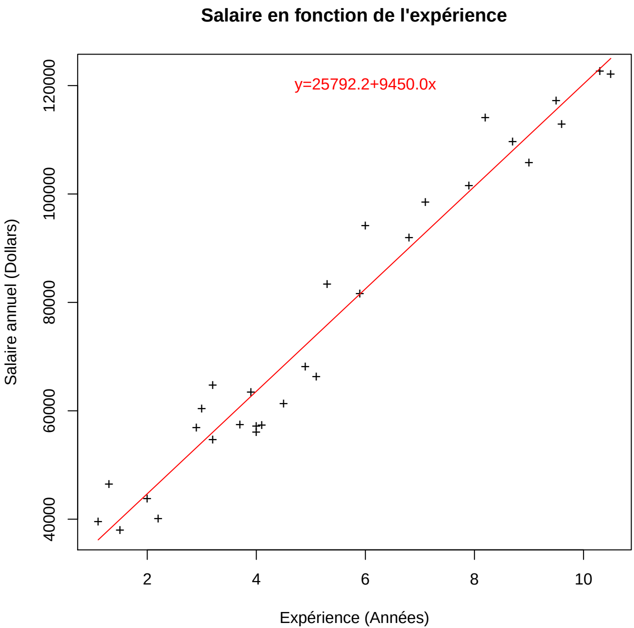
```
#lecture de la table de données, et création d'une variable sal qui est un
> sal=read.table('Salary_Data.csv',sep=',',dec='.',header=TRUE)
#Estimation des paramètres et plus...
> reslm=lm(Salary~YearsExperience,data=sal)
> summary(reslm)


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      25792.2     2273.1   11.35 5.51e-12 ***
YearsExperience   9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared:  0.957,Adjusted R-squared:  0.9554
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

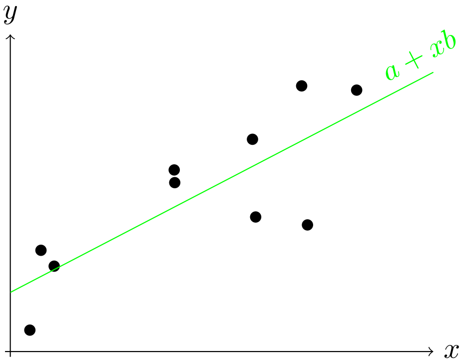
Application au cas des salaires

- $a = 25792.2 \quad \$ =>$ Quelqu'un sans expérience gagne en moyenne 25 792 $ par an...
  Attention : cela n'est vrai que si il existe des données pour lesquelles X est proche de 0. Sinon, on fait de l'extrapolation, on dit quelque chose en dehors du nuage de points et cela devient faux. Ici c'est faux.
- $b = 9450.0 \quad \$/an =>$ Pour chaque année d'expérience gagnée, le salaire annuel augmente en moyenne de 9450 $.
  Attention : Cela n'est vrai que pour une expérience se trouvant ans le nuage de points.
- $s = 5788 \quad \$ =>$ l'écart type des erreurs est estimé à 5 788 $ par an. On retrouve l'estimation de $\sqrt{S^2}$ sous le terme `residual standard error`.

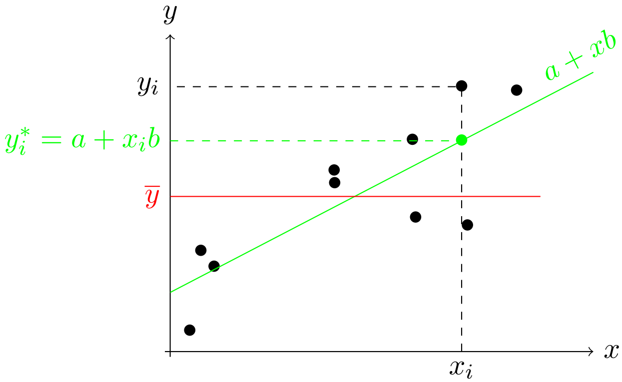


Salaire en fonction de l'expérience
y=25792.2+9450.0x
Salaire annuel (Dollars)
40000
60000
80000
100000
120000
2
4
6
8
10
Expérience (Années)

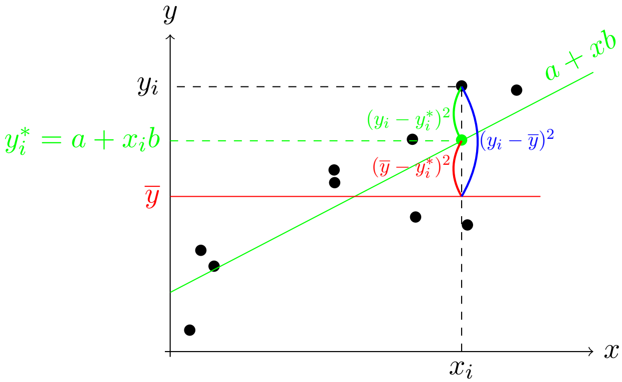
Décomposition de la variance



Décomposition de la variance

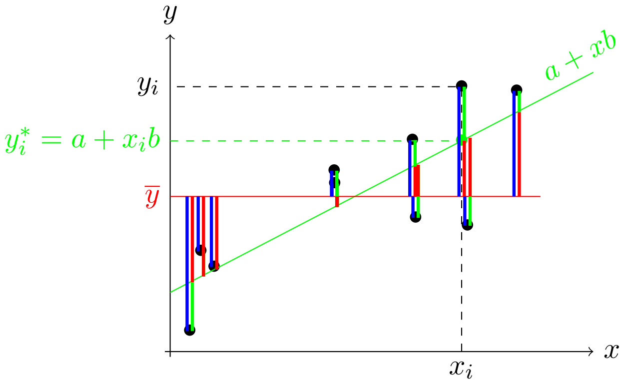


Décomposition de la variance



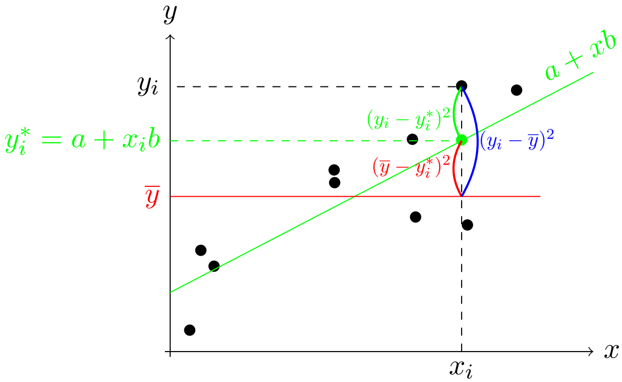
$SCM = \sum_{i=1}^{n}(y_i^* - \overline{y})^2$, somme des carrés du modèle
$SCR = \sum_{i=1}^{n}(y_i - y_i^*)^2$, somme des carrés résiduels
$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$, somme des carrés totale

Décomposition de la variance



$SCM = \sum_{i=1}^{n}(y_i^* - \overline{y})^2$, somme des carrés du modèle
$SCR = \sum_{i=1}^{n}(y_i - y_i^*)^2$, somme des carrés résiduels
$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$, somme des carrés totale

Décomposition de la variance



$SCM = \sum_{i=1}^{n} (y_i^* - \overline{y})^2$, somme des carrés du modèle
$SCR = \sum_{i=1}^{n} (y_i - y_i^*)^2$, somme des carrés résiduels
$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$, somme des carrés totale

Décomposition de la variance

On peut montrer que

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i^*-\overline{y})^2}_{SCM} + \underbrace{\sum_{i=1}^{n}(y_i-y_i^*)^2}_{SCR}$$

On a également les résultats suivants (Cf Th. de Cochran) :

- $\frac{SCM}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sigma^2} \sim \chi^2_{n-2}$
- $\frac{SCR}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sigma^2} \underset{\mathcal{H}_0}{\sim} \chi^2_{1}$
- $\frac{SCT}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \underset{\mathcal{H}_0}{\sim} \chi^2_{n-1}$
- SCR et SCM indépendants.

avec $\mathcal{H}_0$ l'hypothése $\{\beta = 0\}$.

Coefficient de détermination R^2

Définition

On appelle coefficient de détermination de la régression la proportion R^2 définie par

$$R^2 = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SCM}{SCT}$$

- R^2 représente la part de la variabilité des données expliquée par le modèle.
- Plus le R^2 est grand, plus la variable explicative X explique une grande part de la variabilité de variable Y.
- Dans le cadre de la régression simple

$$R^2 = r^2$$

avec r le coefficient de corrélation de Pearson.

Test d'analyse de la variance

- On considère l'hypothése suivante

$$\mathcal{H}_0 = \{\beta = 0\} = \left\{ \begin{array}{c} \text{le modèle de régression linéaire simple} \\ \text{est inutile par rapport à un modèle avec} \\ \text{une moyenne constante} \end{array} \right\}$$

- Sous l'hypothèse $\mathcal{H}_0$, $F = \frac{SCM/1}{SCR/(n-2)} \sim \mathcal{F}_{1,n-2}$.
- On rejette $\mathcal{H}_0$ si F est trop grande, c'est à dire si

$$F_{obs} \geq f_{1,n-2,1-\alpha}$$

- La probabilité critique associée à ce test est :

$$p_c = P(V > F_{obs})$$

avec $V \sim \mathcal{F}_{1,n-2}$.

Application : test d'analyse de la variance

```
> anova(reslm)
Analysis of Variance Table

Response: Salary
                Df     Sum Sq    Mean Sq F value    Pr(>F)
YearsExperience  1 2.0857e+10 2.0857e+10  622.51 < 2.2e-16 ***
Residuals       28 9.3813e+08 3.3505e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `Df` degrés de liberté 1 et $n-2$
- `Sum Sq` Somme des carrés SCM et SCR
- `Mean Sq` Somme des carrées moyens $SCM/1$ et $SCM/(n-2)$
- `F value` Statistique de Fisher du test d'analyse de la variance
- `Pr (>F)` Probabilité critique associée

$=>$ On rejette l'hypothèse , le modèle de régression est utile devant un modèle avec moyenne constante.

Test sur le paramètre α

- On considère les hypothèses $\mathcal{H}_0 = \{\alpha = \alpha_0\}$ et $\mathcal{H}_1 = \{\alpha \neq \alpha_0\}$.
- On rappelle que
  - $A \sim \mathcal{N}(\alpha, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}))$
  - $\frac{\sum_{i=1}^n (Y_i - Y_i^*)^2}{\sigma^2} = \frac{S^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$
  - Les deux estimateurs sont indépendants.
- Sous l'hypothèse $\mathcal{H}_0$, $\frac{(A-\alpha_0)}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim \mathcal{T}_{n-2}$
- On rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$, si $\frac{(A-\alpha_0)}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}}$ s'écarte trop de 0.
- C'est à dire pour un risque de niveau α, si

$$|a - \alpha_0| \geq t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

avec $t_{n-2,1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n-2$ dl.

Test sur le paramètre β

- On considère les hypothèses $\mathcal{H}_0 = \{\beta = \beta_0\}$ et $\mathcal{H}_1 = \{\beta \neq \beta_0\}$.
- On rappelle que
  - $B \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2})$
  - $\frac{\sum_{i=1}^n (Y_i - Y_i^*)^2}{\sigma^2} = \frac{S^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$
  - Les deux estimateurs sont indépendants.
- Sous l'hypothèse $\mathcal{H}_0$, $\frac{(B-\beta_0)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{S} \sim \mathcal{T}_{n-2}$
- On rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$, si $\frac{(B-\beta_0)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{S}$ s'écarte trop de 0.
- C'est à dire, pour un risque de niveau α, si

$$|b - \beta_0| \geq t_{n-2,1-\alpha/2} \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

avec $t_{n-2,1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n-2$ dl.

Intervalle de confiance pour α et β

A partir des lois de A et B énoncées précédemment, on peut obtenir des statistiques pivotales et des intervalles de confiance pour α et β.

- Intervalle de confiance de niveau $1-\gamma$ pour α :

$$IC_{1-\gamma}(\alpha) = \left[a \pm t_{n-2,1-\gamma/2} s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \right]$$

- Intervalle de confiance de niveau $1-\gamma$ pour β :

$$IC_{1-\gamma}(\beta) = \left[b \pm t_{n-2,1-\gamma/2} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \right]$$

Application : test sur les paramètres

```
> summary(reslm)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      25792.2     2273.1   11.35 5.51e-12 ***
YearsExperience   9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Ligne 1 `(Intercept)` α et ligne 2 `YearsExperience` β
- `Estimate` Estimations de α et β
- `Std. Error` Estimation des écarts types de A et B.
- `t value` Statistique du test de Student de nullité de α et β
- `Pr(>|t|)` Probabilités critiques associées

$=>$ α et β sont significativement non nuls.

Intervalle de confiance pour $E(Y_j|X = x_j)$

Comme

- $Y_j^* \sim \mathcal{N}(\alpha + \beta x_j, \sigma^2(\frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}))$
- $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$
- S^2 est indépendant de $\overline{Y}$, B et A, donc de Y_j^*,

on en déduit un intervalle de confiance de niveau $1 - \gamma$ pour $E(Y_j)$

$$IC_{1-\gamma}(E(Y_j)) = \left[y_j^* \pm t_{n-2,1-\gamma/2} s \sqrt{\frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

Remarque : Plus x_i est loin de $\overline{x}$ plus l'intervalle de confiance est grand. On obtient une hyperbole de confiance, lorsque x_i est proche de $\overline{x}$ c'est le terme $1/n$ qui domine, mais si x_i est éloigné de $\overline{x}$ c'est le terme quadratique $(x_i - \overline{x})^2$ qui domine.

Intervalle de prédiction pour une nouvelle donnée

On dispose d'un nouvel x_0 et l'on souhaite **prédire** le Y_0 correspondant avec notre modèle. C'est à dire $Y_0 = \alpha + \beta x_0 + E_0$, avec $E_0 \sim \mathcal{N}(0, \sigma^2)$. On peut prédire Y_0 avec le modèle estimé :

$$Y_0^p = A + Bx_0 + E_0^p \text{ avec } E_0^p \sim \mathcal{N}(0, S^2)$$

- La prédiction est $y_0^p = a + b_{x0}$
- La variance de la prédiction est donnée par

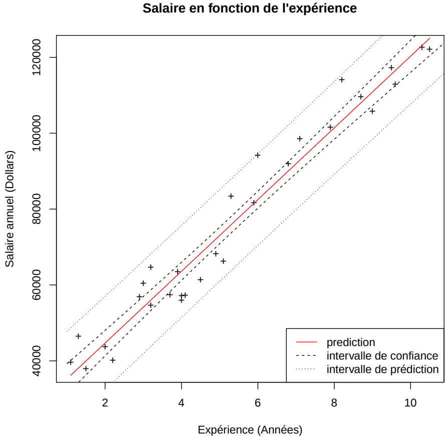
$$V(Y_0^p) = V(A + Bx_0) + V(E_0^p) = S^2 \left(1 + \frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

Remarques :

- la nouvelle donnée donnée n'intervient pas dans le calcul de A et B.
- Deux sources d'incertitudes sur Y_0^p : l'incertitude sur a, b et σ^2 et la variabilité de l'erreur.
- la variance augmente lorsque x_0 s'éloigne de $\overline{x}$.

Application : intervalle de confiance et de prédiction sur Y_0^*

```
> precon=predict(reslm,newdata=data.frame(YearsExperience=seq(1,11,0.01))
+                   ,interval="confidence")
> prepre=predict(reslm,newdata=data.frame(YearsExperience=seq(1,11,0.01))
+                   ,interval="prediction")
```

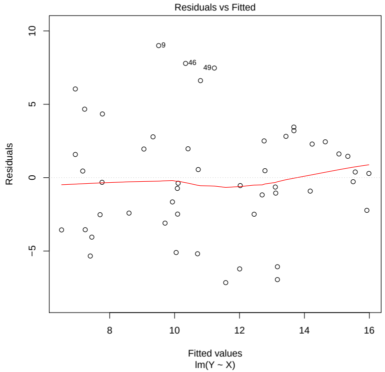


Validation des hypothèses du modèle

1. Hypothèse d'indépendance
2. Hypothèse d'homoscédasticité
3. Hypothèse de normalité
4. Hypothèse de linéarité

=> Cette étape est basée sur l'analyse des résidus.
=> On s'appuie sur des méthodes graphiques.
=>Il n'y a pas de règle fixe pour la prise de décision.

Graphique des résidus vs valeurs prédites

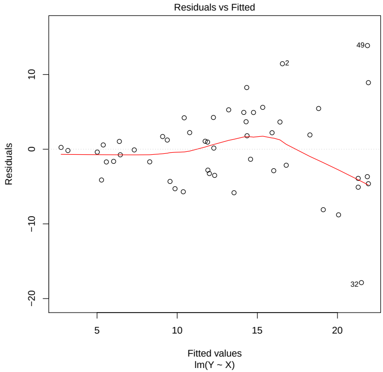


=> Correct

On vérifie

- Absence de forme dans le nuage de point
  - forme en entonnoir : hétéroscédasticité
  - autre forme : non linéarité
- Absence d'individu abberant
  Une donnée (x_i, y_i) pour laquelle le résidu en valeur absolue est très grand ($\approx$ plus de $2s$.

Graphique des résidus vs valeurs prédites

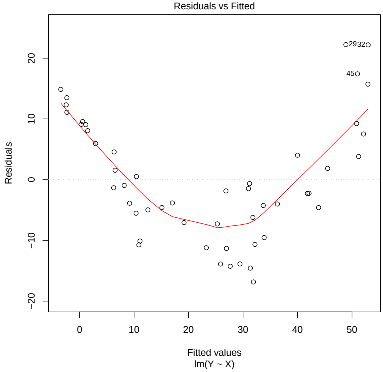


=> Forme en entonnoir

On vérifie

- Absence de forme dans le nuage de point
  - forme en entonnoir : hétéroscédasticité
  - autre forme : non linéarité
- Absence d'individu abberant
  Une donnée (x_i, y_i) pour laquelle le résidu en valeur absolue est très grand ($\approx$ plus de $2s$.

Graphique des résidus vs valeurs prédites

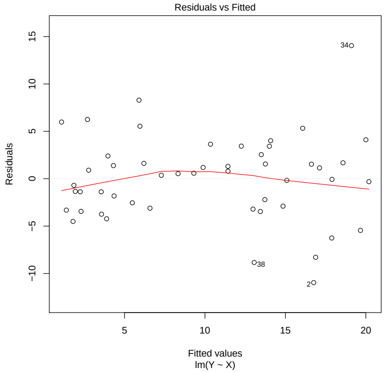


$=>$ Forme quadratique

On vérifie

- Absence de forme dans le nuage de point
  - forme en entonnoir : hétéroscédasticité
  - autre forme : non linéarité
- Absence d'individu abberant
  Une donnée (x_i, y_i) pour laquelle le résidu en valeur absolue est très grand ($\approx$ plus de $2s$.

Graphique des résidus vs valeurs prédites

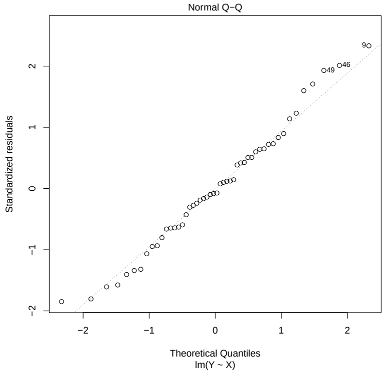


=> Individu 34 aberrant

On vérifie

- Absence de forme dans le nuage de point
  - forme en entonnoir : hétéroscédasticité
  - autre forme : non linéarité
- Absence d'individu abberant
  Une donnée (x_i, y_i) pour laquelle le résidu en valeur absolue est très grand ($\approx$ plus de $2s$.

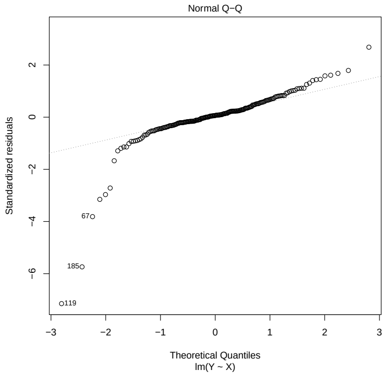
QQ plot



=> Correct

- On vérifie la normalité.
- On compare :
  - les quantiles des résidus estimés
  - avec l'espérance des même quantiles sous l'espérance de normalité.
- Si l'hypothèse de normalité est vérifiée les points doivent être à peu près alignés sur la première bissectrice.

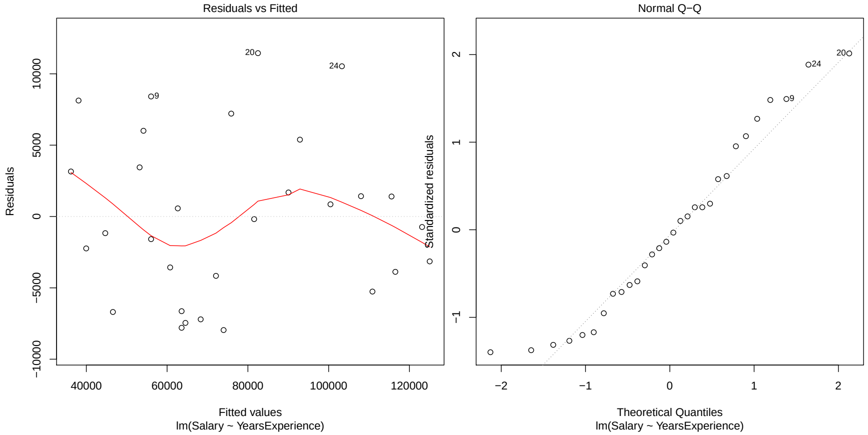
QQ plot



=> Queues de distributions plus lourdes que la loi normale

- On vérifie la normalité.
- On compare :
  - les quantiles des résidus estimés
  - avec l'espérance des même quantiles sous l'espérance de normalité.
- Si l'hypothèse de normalité est vérifiée les points doivent être à peu près alignés sur la première bissectrice.

Application : validation des hypothèses du modèle



Écriture du modèle de régression simple sous forme matricielle

Nous avons écrit le modèle ainsi :

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$

c'est à dire

$$\begin{cases} Y_1 & \sim & \alpha + \beta x_1 & + & E_1 \\ Y_2 & \sim & \alpha + \beta x_2 & + & E_2 \\ & & \dots & & \\ Y_n & \sim & \alpha + \beta x_n & + & E_n \end{cases}$$

avec $E_i \sim \mathcal{N}(0, \sigma^2), iid.$

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}}_{\mathbf{E}}$$

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

- $\mathbf{Y}$ et $\mathbf{E}$ sont des vecteurs gaussiens.
- $\boldsymbol{\beta}$ est le vecteur des paramètres de régression à estimer.
- $\mathbf{X}$ est la matrice de design.

De manière concise on écrira

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \text{ avec } \mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$$