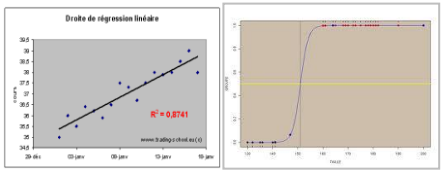
I. MACHINE LEARNING

DATA MINING : MÉTHODES PRÉDICTIVES

RÉGRESSION

LINÉAIRE / LOGISTIQUE



Mohamed Heny SELMI

medheny.selmi@esprit.tn

DEUX FAMILLES DE TECHNIQUES

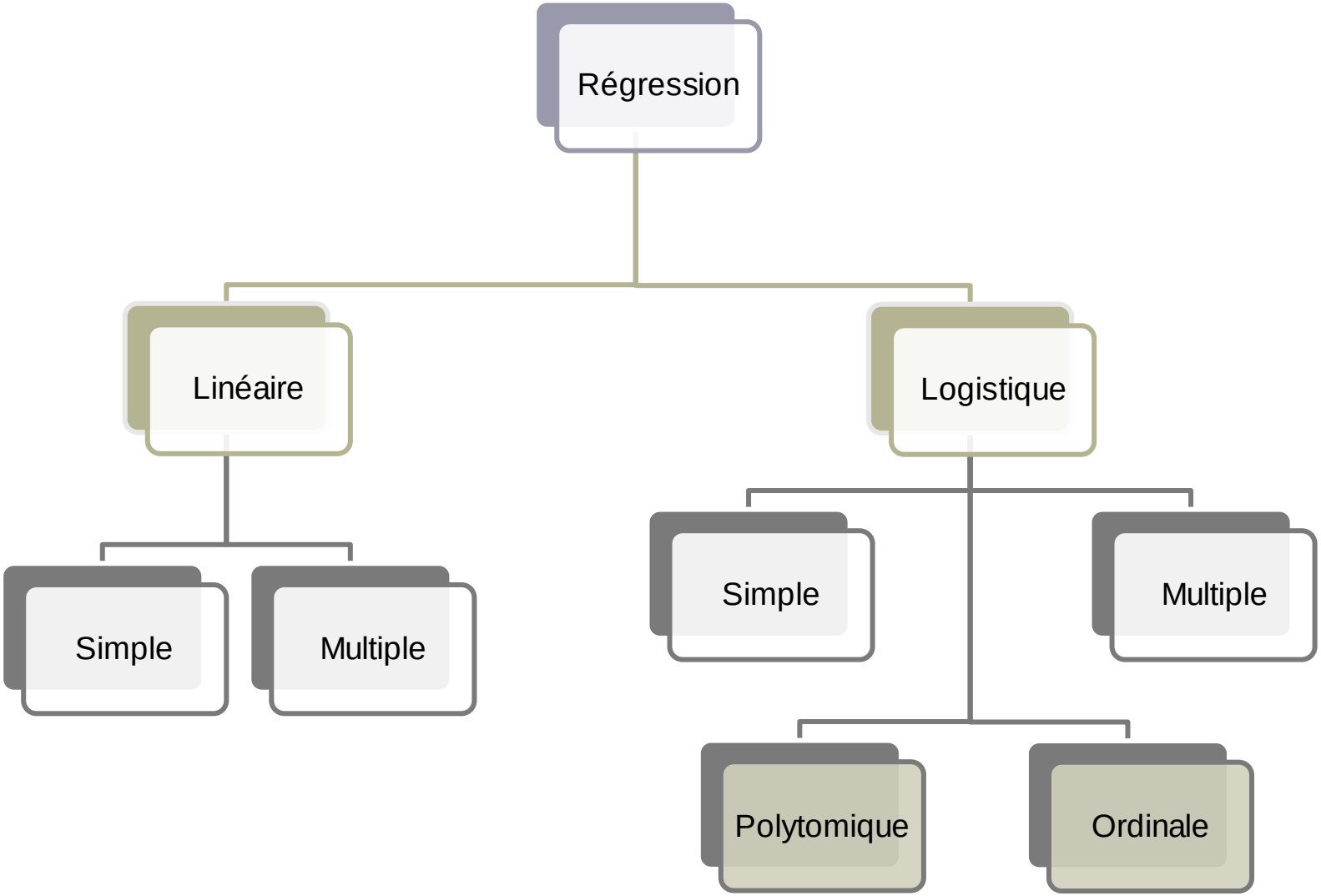
Méthodes Descriptives

- Analyse en Composantes Principales ACP
- Méthodes des Centres Mobiles K-Means
- Classification Ascendante Hiérarchique CAH

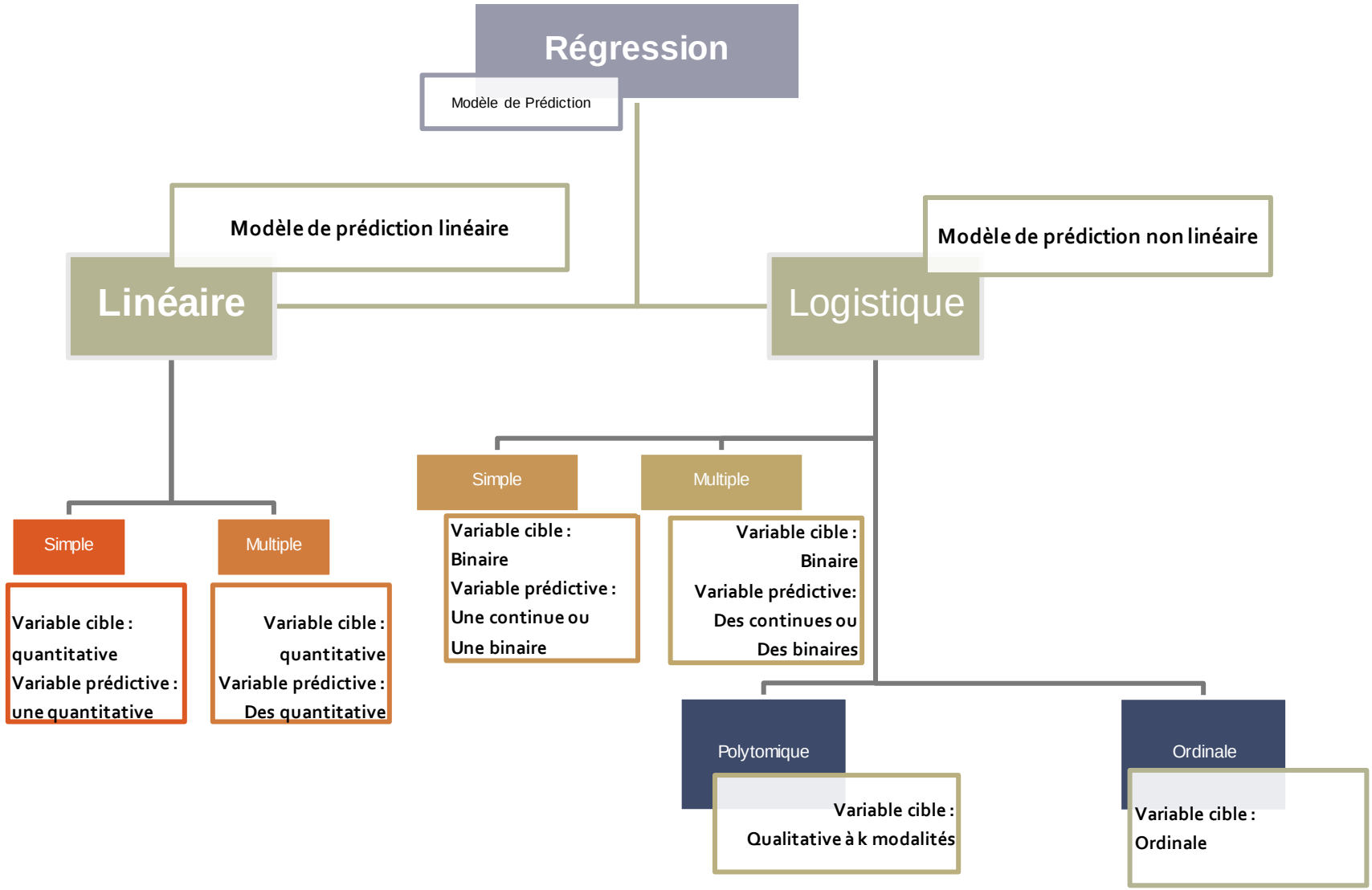
Méthodes Prédictives

- Arbres de Décisions
- **Régression Linéaire, Logistique**
- **Analyse Discriminante**
- **Réseaux de Neurones**

TYPES DE RÉGRESSION



TYPES DE RÉGRESSION



PRINCIPE DE LA RÉGRESSION

- L'analyse de la régression permet d'étudier le type de relation pouvant exister entre une certaine variable (dépendante) dont on veut expliquer les valeurs et une ou plusieurs autres variables qui servent à cette explication (variables indépendantes)
- En d'autres termes, l'analyse de la régression permet d'étudier les variations de la variable dépendante en fonction des variations connues des variables indépendantes.
- De détecter les individus atypiques

RÉGRESSION LINÉAIRE

OBJECTIFS DE LA RÉGRESSION LINÉAIRE

- ✓ Le modèle de prédiction ***LINEAIRE*** consiste à estimer la valeur d'une variable **continue** (dite « à expliquer »,« cible », en fonction de la valeur **d'un certain nombre d'autres variables** (dites « explicatives », « de contrôle », ou « indépendantes »)
- ✓ Cette variable « cible » peut être par exemple :
  - le poids : en fonction de la taille
  - le prix d'un appartement : en fonction de sa superficie, de l'étage et du quartier
  - la consommation d'électricité : en fonction de la température extérieure et de l'épaisseur de l'isolation

BESOIN DE LA RÉGRESSION LINÉAIRE

- Pour estimer la relation entre une **variable dépendante (*Y*) quantitative** et plusieurs variables **indépendantes (*X_1*, *X_2*, ...)**
- Un modèle de régression d'une variable expliquée sur une ou plusieurs variables explicatives dans lequel on fait l'hypothèse que la fonction qui relie les variables explicatives à la variable expliquée est linéaire selon un ensemble de paramètres.
- Dans ce modèle linéaire simple : X et Y deux variables continues

✓ Les valeurs x_i de X sont contrôlées et sans erreur de mesure

✓ On observe les valeurs correspondantes $y_1, \ldots, y_n$ de Y

Exemples :

✓ *X peut être le temps et Y une grandeur mesurée à différentes dates*

✓ *Y peut être la différence de potentiel mesurée aux bornes d'une résistance pour différentes valeurs de l'intensité X du courant*

EXEMPLE DE RÉGRESSION LINÉAIRE

$$Y = f(X_1, X_2, X_3, \ldots, X_n)$$

Coût du loyer

Nombre de pièces

Services offerts
(piscine, stationnement intérieur, etc.)

L'étage dans l'immeuble

...

Estimer le coût du loyer en fonction :

du nombre de pièces,

du niveau d'étage dans l'immeuble,

des services offerts ...

RÉGRESSION LINAIRE MULTIPLE

Equation de régression multiple :

Cette équation précise la façon dont la variable dépendante est reliée aux variables explicatives :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Où $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ sont les paramètres et ε est un bruit aléatoire représentant le terme d'erreur.

PS: on suppose l'indépendance linéaire des X_i

LES TERMES DE L'ÉQUATION

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots \beta_p x_{pi} + \varepsilon_i$$

$i^{\text{ème}}$ observation de Y

Terme constant

Influence de la variable X_1

Influence de la variable X_p

Résidu de la $i^{\text{ème}}$ observation

FORME MATRICIELLE

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

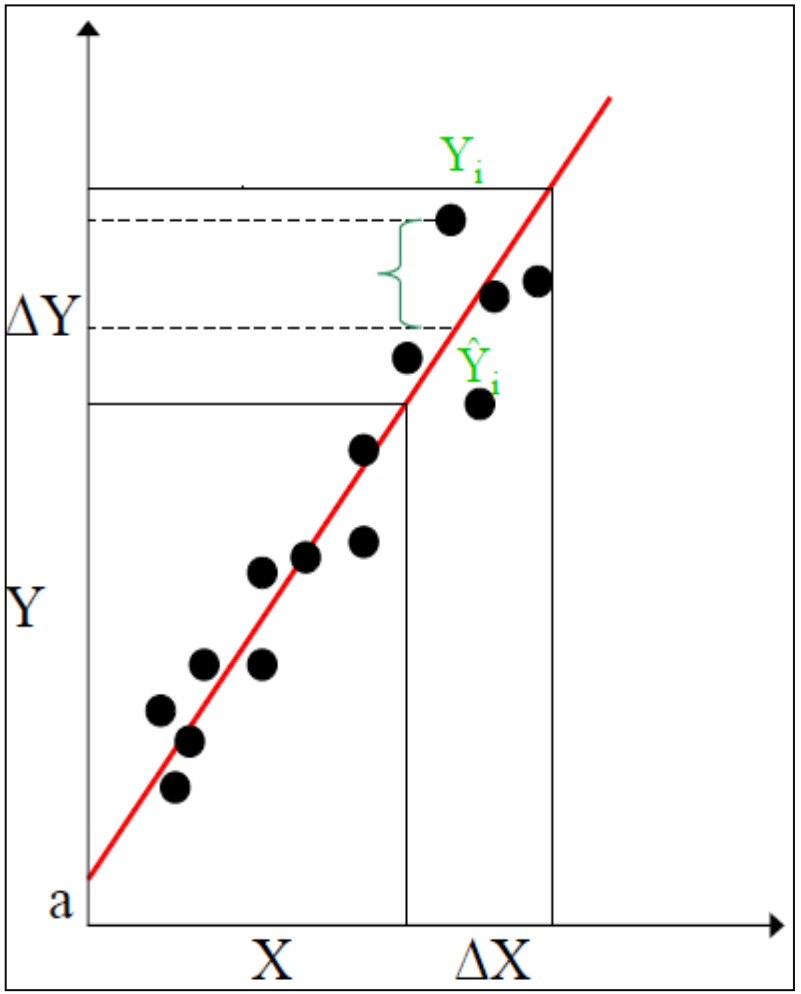
PROCESSUS D'ESTIMATION
MÉTHODE DES MOINDRES CARRÉS

Estimation des coefficients de régression / méthode des moindres carrés ordinaires :

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$$

Le principe de l'estimation des coefficients de régression :

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



CALCUL DES COEFFICIENTS ESTIMATEURS

La méthode des moindres carrés donne pour résultat :

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

Interprétation des coefficients de régression estimés

- La pente $\hat{\beta}_k$ ($k \neq 0$)

L'estimée de Y varie d'un facteur égal à $\hat{\beta}_k$ lorsque X_k augmente d'une unité, les autres variables étant maintenues constantes.

- L'ordonnée à l'origine $\hat{\beta}_0$:

C'est la valeur moyenne de Y lorsque toutes les X_i sont nulles.

COEFFICIENTS DE RÉGRESSION ET TESTS

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	1467,643	62,422		23,512	,000
	TEMPERAT	-37,060	2,295	-,866	-16,147	,000
	ISOLATIO	-29,774	3,492	-,457	-8,526	,000

a. Variable dépendante : CONSOMMA

Valeur des coefficients

Écart-type des estimateurs

Coefficients comparables entre eux

Statistique t de Student

Une valeur t > 2 ou t < - 2 est significative à 95 % d'un coeff ≠ 0

COEFFICIENTS DE RÉGRESSION ET TESTS

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	1467,643	62,422		23,512	,000
	TEMPERAT	-37,060	2,295	-,866	-16,147	,000
	ISOLATIO	-29,774	3,492	-,457	-8,526	,000

a. Variable dépendante : CONSOMMA

Valeur des coefficients

Écart-type des estimateurs

Coefficients comparables entre eux

Statistique t de Student

Une valeur t > 2 ou t < - 2 est significative à 95 % d'un coeff ≠ 0

L'APPORT MARGINAL DE X_J EST-IL SIGNIFICATIF ?

Modèle : $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots + \beta_k X_k + \varepsilon$

Test : $H_0 : \beta_j = 0$ (*On peut supprimer* X_j)

$H_1 : \beta_j \neq 0$ (*Il faut conserver* X_j)

Exemple : indiquer les variables significatives du problème !

SÉLECTION DES VARIABLES
RÉGRESSION PAS À PAS DESCENDANTE (BACKWARD)

i. On part du modèle complet.

ii. A chaque étape, on enlève la variable X_j ayant l'apport marginal le plus faible à condition que cet apport soit non significatif

EXEMPLE : CAS DE VENTES SEMESTRIELLES

Variable à expliquer :

Y = Ventes semestrielles

Variables explicatives :

X_1 = Marché total

X_2 = Remises aux grossistes

X_3 = Prix

X_4 = Budget de Recherche

X_5 = Investissement

X_6 = Publicité

X_7 = Frais de ventes

X_8 = Total budget publicité de la branche

PREMIÈRE ÉTAPE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.898[a]	.806	.752	256.29

a. Predictors: (Constant), Total publicité de la branche, Marché total, Remises aux grossistes, Budget de recherche, Investissements, Publicité, Prix, Frais de ventes

TPUB = Total budget publicité de la branche

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3129.231	641.355	4.879	.000
	MT	4.423	1.588	2.785	.009
	RG	1.676	3.291	.509	.614
	PRIX	-13.526	8.305	-1.629	.114
	BR	-3.410	6.569	-.519	.608
	INV	1.924	.778	2.474	.019
	PUB	8.547	1.826	4.679	.000
	FV	1.497	2.771	.540	.593
	TPUB	-2.15E-02	.401	-.054	.958

a. Dependent Variable: VENTES

DEUXIÈME ETAPE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.898[a]	.806	.760	251.99

a. Predictors: (Constant), Frais de ventes, Remises aux grossistes, Publicité, Investissements, Budget de recherche, Prix, Marché total

b. Dependent Variable: Ventes

BR = Budget de Recherche

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3115.648	579.517	5.376	.000
	MT	4.426	1.561	2.836	.008
	RG	1.706	3.191	.535	.597
	PRIX	-13.445	8.029	-1.675	.104
	BR	-3.392	6.451	-.526	.603
	INV	1.931	.756	2.554	.016
	PUB	8.558	1.784	4.798	.000
	FV	1.482	2.710	.547	.588

a. Dependent Variable: VENTES

TROISIÈME ÉTAPE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.897[a]	.804	.766	249.04

a. Predictors: (Constant), Frais de ventes, Remises aux grossistes, Publicité, Investissements, Prix, Marché total

b. Dependent Variable: Ventes

FV = Frais de ventes

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3137.547	571.233	5.493	.000
	MT	4.756	1.412	3.368	.002
	RG	1.705	3.153	.541	.593
	PRIX	-14.790	7.521	-1.966	.058
	INV	1.885	.742	2.539	.016
	PUB	8.519	1.761	4.837	.000
	FV	.950	2.484	.382	.705

a. Dependent Variable: VENTES

QUATRIÈME ÉTAPE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896[a]	.803	.772	245.69

a. Predictors: (Constant), Publicité, Remises aux grossistes, Marché total, Investissements, Prix

b. Dependent Variable: Ventes

RG = Remises aux grossistes

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3084.009	546.374	5.645	.000
	MT	5.222	.704	7.415	.000
	RG	1.700	3.111	.546	.589
	PRIX	-13.467	6.589	-2.044	.049
	INV	1.984	.686	2.893	.007
	PUB	8.328	1.666	4.998	.000

a. Dependent Variable: VENTES

CONDITION D'ARRÊT

Toutes les « Signification » < 0.05

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3084.009	546.374	5.645	.000
	MT	5.222	.704	7.415	.000
	RG	1.700	3.111	.546	.589
	PRIX	-13.467	6.589	-2.044	.049
	INV	1.984	.686	2.893	.007
	PUB	8.328	1.666	4.998	.000

a. Dependent Variable: VENTES

RÉGRESSION LOGISTIQUE

GÉNÉRALITÉ

Y variable cible binaire Y = 0 / 1

Xj p variables explicatives continues, binaires ou qualitatives :

- p = 1 régression logistique simple

- p > 1 régression logistique multiple

Généralisation : régression logistique polytomique

- la variable cible Y est qualitative à k modalités

- cas particulier : Y ordinale (régression logistique ordinale)

Pb de régression : modéliser l'espérance conditionnelle

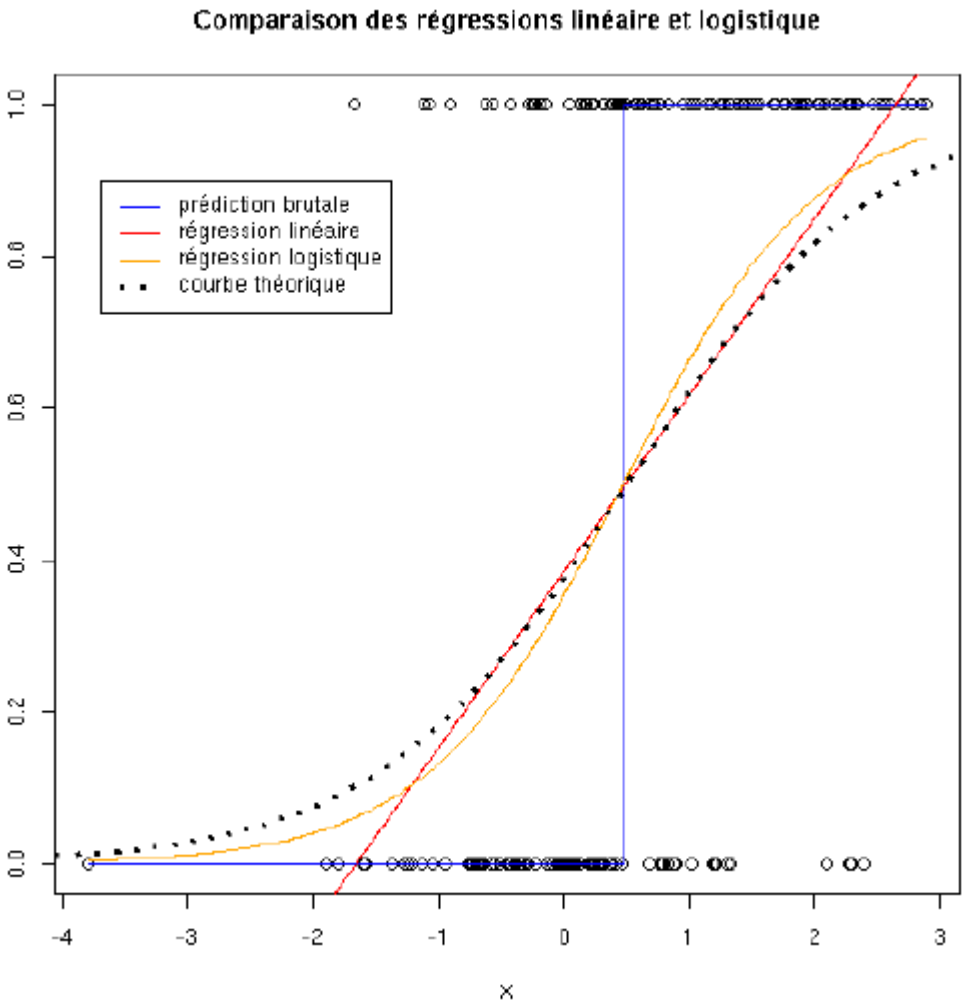
$$E(Y/X=x) = Prob(Y=1/X=x)$$

sous la forme $E(Y/X=x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

PRÉDICTION D'UNE VARIABLE BINAIRE

Visiblement la régression linéaire ne convient pas (distribution des résidus !)

La régression LOGISTIQUE décrit mieux l'aspect comportemental des points individuels

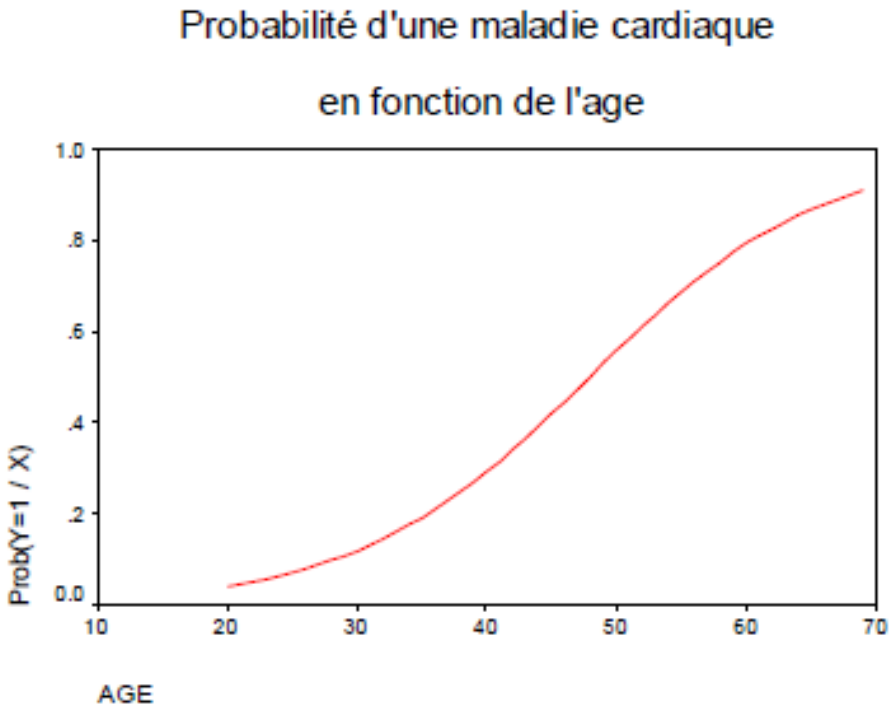


PRÉDICTION D'UNE VARIABLE BINAIRE

- La figure fait pressentir que ce n'est pas une fonction linéaire de :

$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ qu'il faut appliquer, mais une courbe en S

- Les courbes en S sont courantes en biologie et en épidémiologie



FONCTION DE LIEN

On écrit donc π(x) = Prob(Y=1/X=x) sous la forme :

$$\pi(x) = \frac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}}$$

$$\Leftrightarrow \quad Log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Fonction de lien : Logit(π(x))

ESTIMATION DES PARAMÈTRES (BINAIRE)

Les données

vecteur X	Y
x^1	y^1
⋮	⋮
x^i	y^i
⋮	⋮
x^n	y^n

$y^i = 0$ ou 1

Le modèle

$$\pi(x^i) = P(Y = 1 / X = x^i)$$

$$= \frac{e^{\beta_0 + \sum_j \beta_j x^i{}_j}}{1 + e^{\beta_0 + \sum_j \beta_j x^i{}_j}}$$