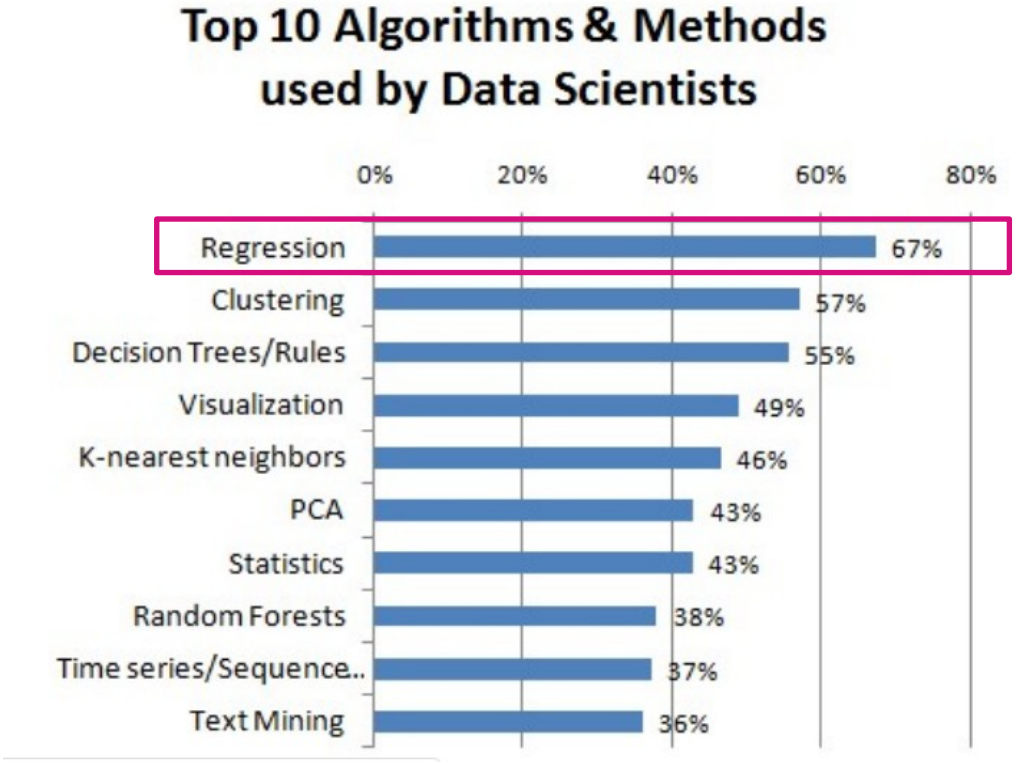
HEC MONTRÉAL

Analyse multidimensionnelle appliquée

Régression logistique

Contenu

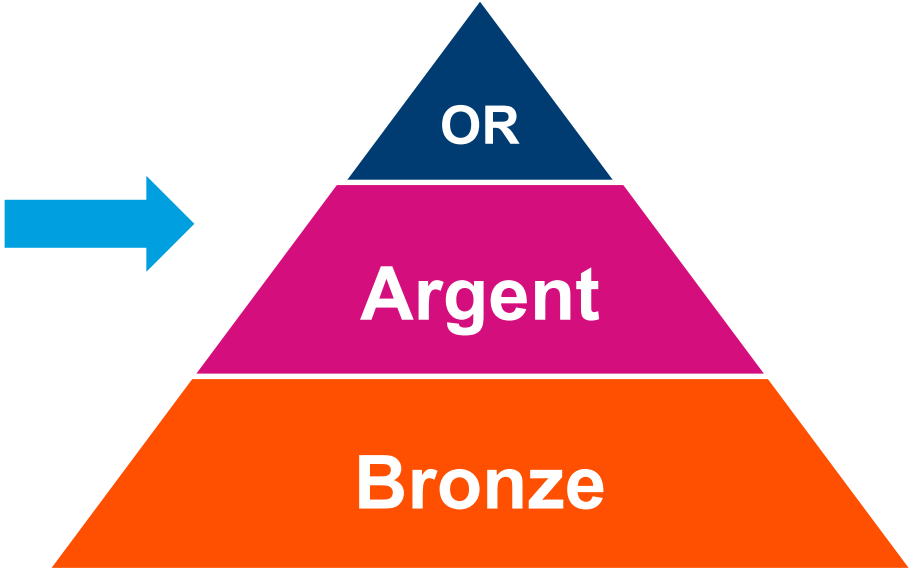
- Mise en contexte
- Modèle de régression logistique
- Interprétation des paramètres
- Tests d'hypothèses
- Calcul de prévision



HEC MONTRÉAL

4.1 Introduction

Mise en contexte : L'exemple d'une pharmacie



HEC MONTRÉAL

4.1 Introduction

Post-campagne sur les cosmétiques

Envoi d'un courriel aux clients OR

« 20% de rabais durant 3 jours sur les cosmétiques »

N=6000 clients ciblés

Campagne

2500 clients viennent acheter des cosmétiques et bénéficient du rabais de 20%.

Post-campagne

Quelles sont les caractéristiques qui distinguent les clients qui ont bénéficié de l'offre de ceux qui n'ont pas bénéficié de l'offre?

Explicatif? **Prédictif?**

HEC MONTRÉAL

4.1 Introduction

Déterminer la meilleure offre pour un client

Offres promotionnelles

Le client recevra une offre parmi 3 offres possibles.

N=6000 clients ciblés

Probabilité de profiter de l'offre

Calculer la probabilité pour chaque client de profiter de chacune des 3 offres.

Campagne

Envoi d'un courriel avec l'offre la plus pertinente pour chacun des clients.

Produits pour bébé

Cosmétiques

Produits nettoyants

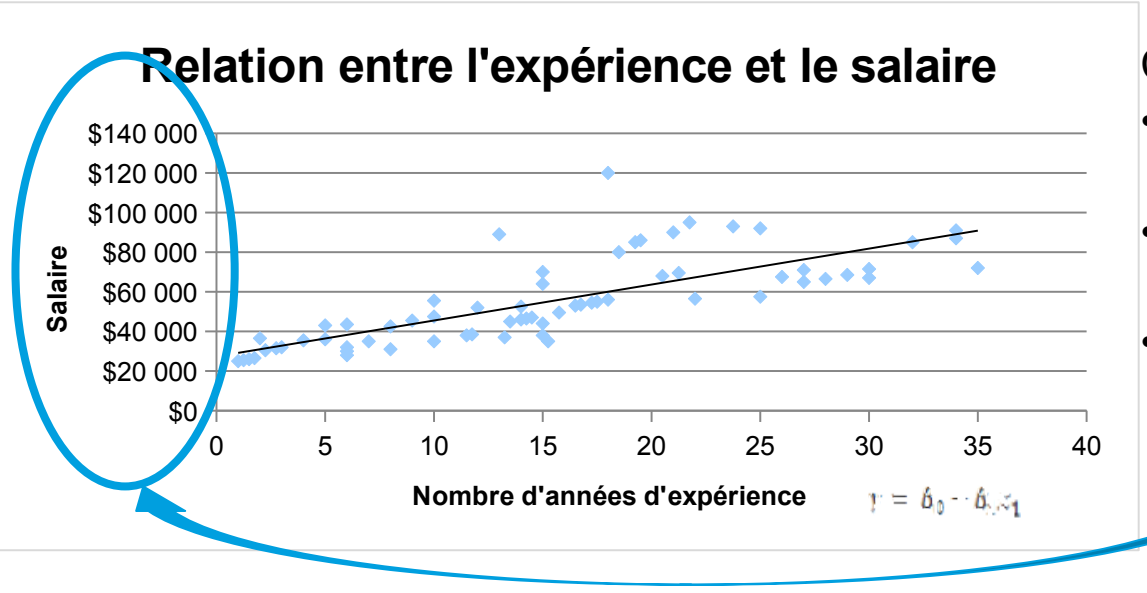
Explicatif?

Prédictif?

HEC MONTRÉAL

4.2 Modèle de régression logistique

Rappel : Régression linéaire



Caractéristiques :

- La formule de la régression est l'équation d'une droite linéaire.
- Cette droite varie entre -l'infini et +l'infini
- Cette méthode est appropriée pour modéliser une variable sur une échelle continue.

Que se passe-t-il si notre cible est plutôt une variable binaire (oui / non)?

HEC MONTRÉAL

4.2 Modèle de régression logistique

Comment modéliser une variable cible binaire?

Cible

ID	Sexe	Age	Intérêt
1	Femme	35	oui
2	Femme	43	oui
3	Homme	28	non
4	Homme	62	non
5	Homme	46	oui

Probabilité de quitter
40%
60%
30%
70%
40%

La probabilité (*p*) varie entre 0 et 1

- Le but est de pouvoir modéliser p en fonction de plusieurs variables indépendantes X_i
- Cependant, *p* varie entre 0 et 1.

HEC MONTRÉAL

4.4 Exemple du PRCA

- Choix de la catégorie de référence :

X1:Quel genre d'emploi occupez-vous?	Proportion de y=1
1=à la maison	59%
2=employé	45%
3=ventes/services	51%
4=professionnel	33%
5=agriculture/ferme	56%

```
proc logistic data=multi.logit1 ;
class x1(ref=last) / param=ref;
model y(ref='0') = x1  / clparm=pl clodds=pl expb;
run;
```

La catégorie de référence est
5= agriculture/ferme

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	0.2549	0.2393	1.1347	0.2868	1.290
x1	1	1	0.0899	0.3972	0.0513	0.8208	1.094
x1	2	1	-0.4406	0.3076	2.0518	0.1520	0.644
x1	3	1	-0.2226	0.2992	0.5536	0.4568	0.800
x1	4	1	-0.9480	0.2934	10.4366	0.0012	0.388

4.4 Exemple du PRCA

- Choix de la catégorie de référence :

X1:Quel genre d'emploi occupez-vous?	Proportion de y=1
1=à la maison	59%
2=employé	45%
3=ventes/services	51%
4=professionnel	33%
5=agriculture/ferme	56%

```
proc logistic data=multi.logit1 ;
class x1(ref='4') / param=ref;
model y(ref='0') = x1  / clparm=pl clodds=pl expb;
run;
```

La catégorie de référence est
4= professionnel

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-0.6931	0.1698	16.6523	<.0001	0.500
x1	1	1	1.0379	0.3596	8.3291	0.0039	2.823
x1	2	1	0.5073	0.2573	3.8881	0.0486	1.661
x1	3	1	0.7253	0.2472	8.6088	0.0033	2.065
x1	5	1	0.9480	0.2934	10.4366	0.0012	2.580

4.4.5 Test du rapport de vraisemblance

Modèle complet : X1 X2 X3 X5 X5 **X6**

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	691.270	544.196
SC	695.485	603.201
-2 Log L	689.270	516.196

Modèle partiel : X1 X2 X3 X5 X5

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	691.270	590.447
SC	695.485	641.022
-2 Log L	689.270	566.447

566.447 – 516.196 = 49.487

Nombre de degrés de liberté = 2

- Il s'agit du nombre de paramètres de plus qui sont estimés dans le modèle complet par rapport au modèle partiel.

HEC MONTRÉAL

4.5 Classification (prévision)

- Pour classifier des observations, il suffit de choisir un point de coupure (souvent 0,5 mais pas toujours).

y	Response Value	Estimated Probability
0	1	0.4008266813
0	1	0.5077519353
0	1	0.4008266813
1	1	0.5077519353
1	1	0.5077519353
0	1	0.4008266813
0	1	0.4008266813
0	1	0.5077519353
1	1	0.5077519353
0	1	0.5077519353
1	1	0.5077519353
0	1	0.4008266813
1	1	0.5077519353
1	1	0.4008266813
0	1	0.5077519353

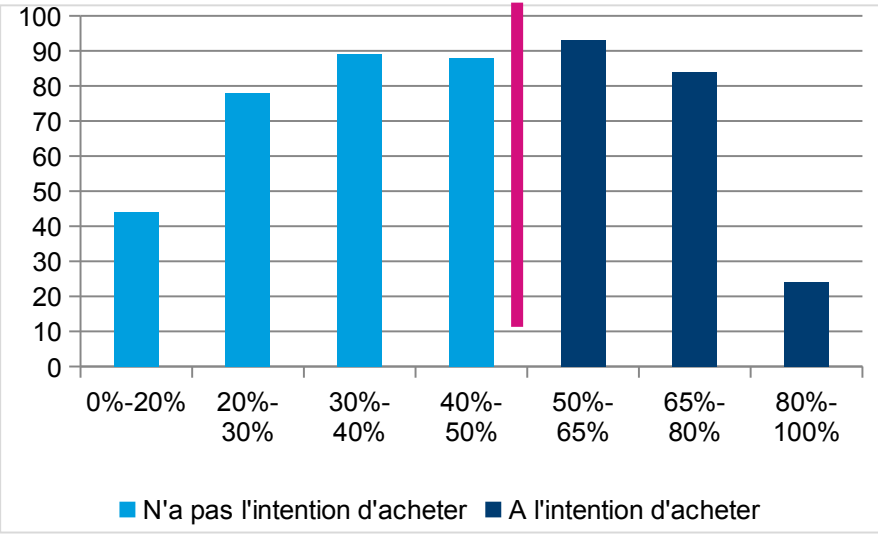
$\hat{y} = 1 \quad car \quad \hat{p} > 0,5$

$\hat{y} = 0 \quad car \quad \hat{p} < 0,5$

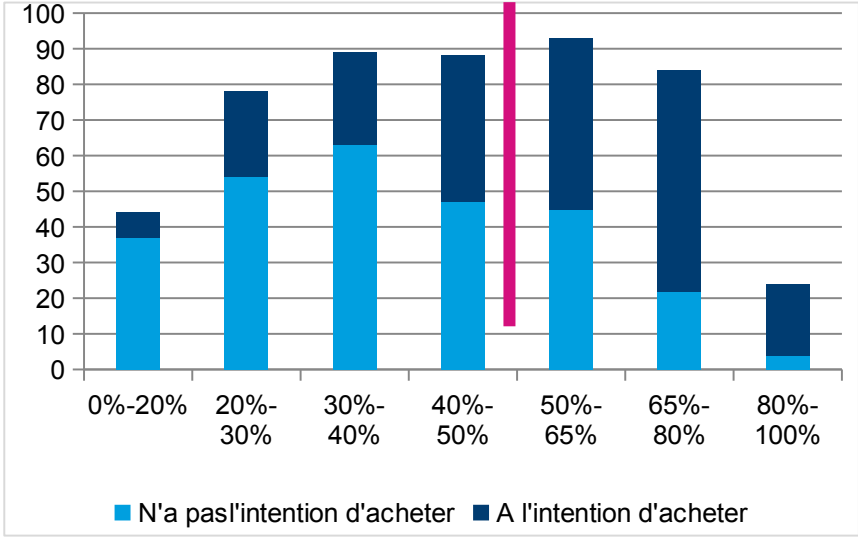
4.5 Classification (prévision)

- Taux de bons classements

Discrimination parfaite (impossible en pratique)



Situation observée



HEC MONTRÉAL