Le modèle linéaire généralisé (logit,...)

Master 2 ISEFAR Classification

Ana Karina Fermin

Université Paris Nanterre

http://fermin.perso.math.cnrs.fr/

Objectif.

On souhaite "expliquer" une variable réponse Y par une variable explicative X (ou plusieurs variables explicatives $X_1, X_2, \ldots, X_p$) lorsque Y est 0 (échec) ou 1 (succès).

Modélisation (cas multiple avec p variables)

- La loi de Y est déterminée par

$$\pi(X) = P(Y = 1|X_1, X_2, \ldots, X_p)$$

- Nous supposons $\pi(X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$, où F est une fonction de répartition inversible donnée avec $\beta_0, \beta_1, \ldots, \beta_p$ inconnus.

Estimation

- En pratique, les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ doivent être déterminés à l'aide des données.
- On utilise la méthode du Maximum de Vraisemblance (MV).
- En général la méthode de MV fournit des estimateurs avec des bonnes propriétés statistiques.

Commençons par définir la fonction log-vraisemblance associée au modèle logit et probit

log-Vraisemblance

$$\mathrm{LV}(\beta) = \sum_{i=1}^{n} Y_i \log(F(X_i)) + (1 - Y_i) \log(1 - F(X_i))$$

avec $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.

Les logiciels de statistiques calculent la fonction $\mathrm{LV}(\beta)$ et cherchent les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ que maximisent cette fonction à l'aide d'un algorithme itérative.

Notre objectif est modéliser

$$\pi(X) = P(Y = 1|X_1, X_2, \ldots, X_p)$$

Nous supposons $\pi(x) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$

Exemples de fonctions F :

- logit : F est la fonction de répartition de la loi logistique.
- probit : F est la fonction de répartition de la loi Gaussienne standard.

Régression logistique

Fonction de répartition de la loi logistique

On parle de régression logit ou logistique lorsque pour tout $t \in \mathbb{R}$,

$$F(t) = \frac{\exp(t)}{1+\exp(t)}.$$

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$

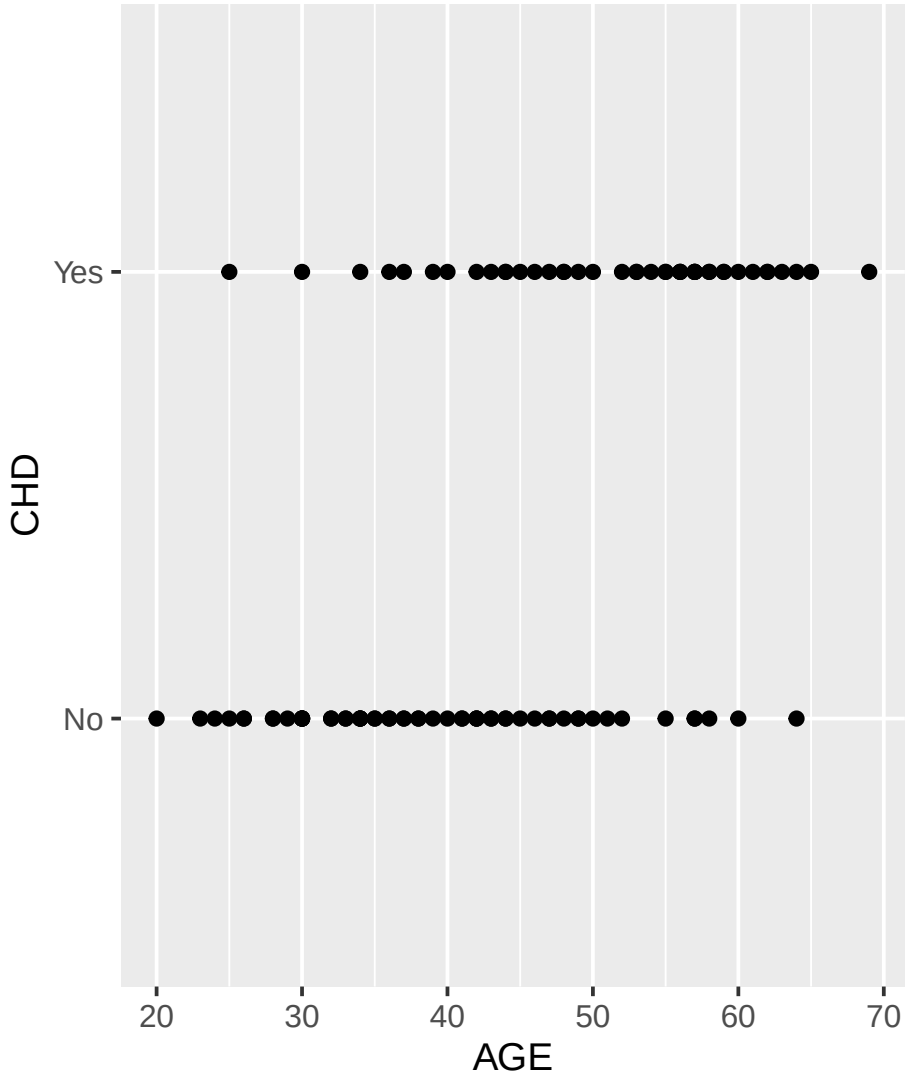
$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Exemple 1 (cf. Ricco Rakotomalala)

On étudie la variable binaire CHD qui prend la valeur 1 si présence d'un problème cardiaque et 0 si absence. On souhait étudier la relation entre CHD et la variable explicative âge (AGE)

Le fichier maladie_cardiovasculaire.txt comporte 100 lignes, dont les cinq premières sont :

```
> head(maladie,5)
  ID AGRP AGE CHD
1  1    1  20   0
2  2    1  23   0
3  3    1  24   0
4  4    1  25   0
5  5    1  25   1
```


Yes
No
CHD
20
30
40
50
60
70
AGE

Données groupées

Supposons que l'on ait K groupes, i.e. seulement K valeurs possibles pour la de variable explicative X, et que pour chaque groupe k, $k = 1, \ldots, K$, on dispose de n_k observations. Ainsi,

$$P(Y_{kj} = 1 | X_k = x_k) = \pi(x_k),\ j \in \{1, \ldots, n_k\}.$$

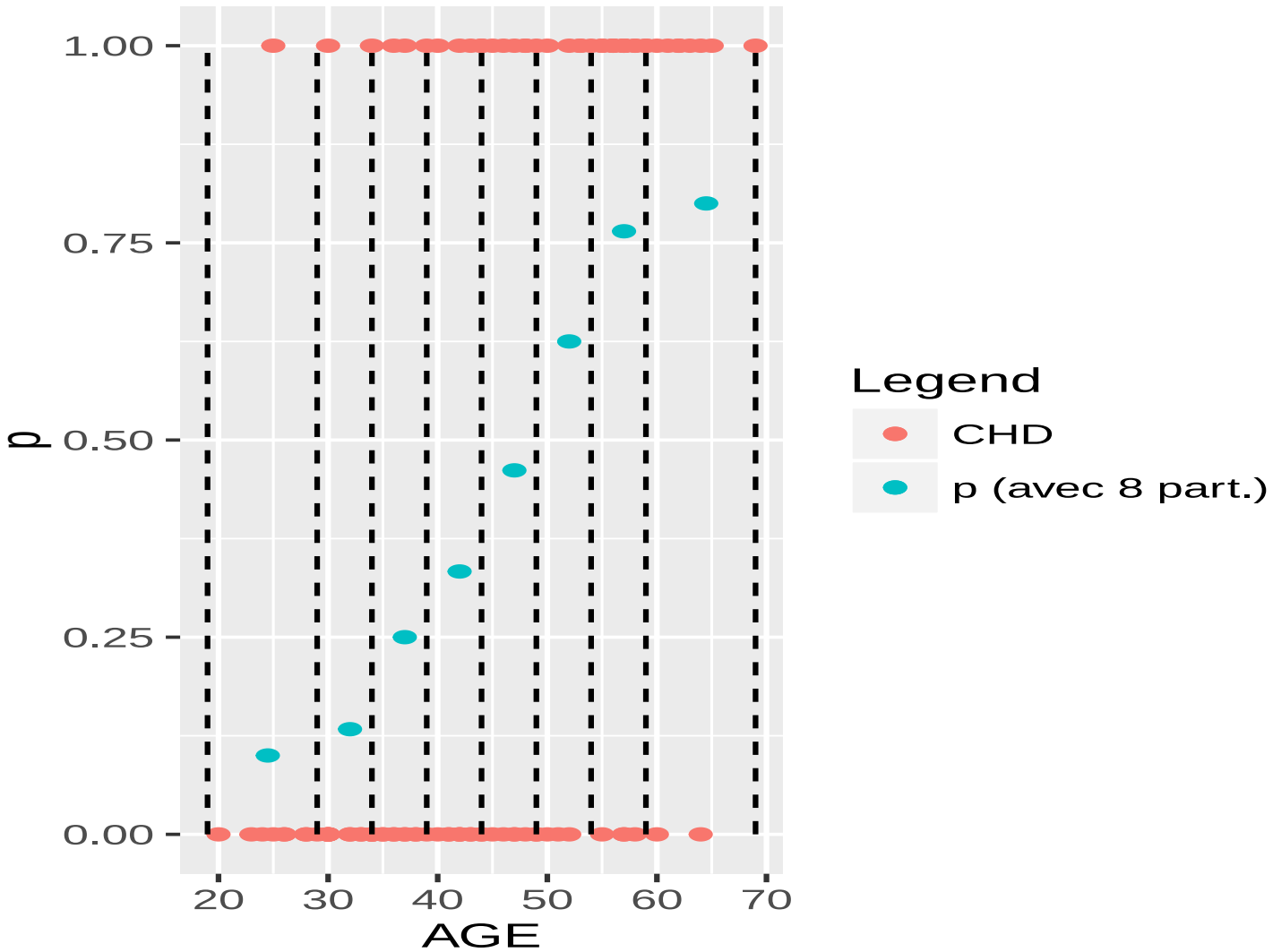
On dit dans ce cas que les données sont groupées. Sinon, on dit que les données sont individuelles

Remarque : On peut ramener des données individuelles au cas de données groupées en segmentant selon les variables explicatives.

Retour à l'exemple 2

Le tableau suivant donne c_k le centre de chaque classe d'age, n_k le nombre de patients selon la classe d'age, la proportion de malades selon la classe d'age $\pi_k = n_k[\mathrm{CHD} = 1]/n_k$, ...

Age_k	c_k	n_k	n_k[CHD=0]	n_k[CHD=1]	π_k
[20,29]	24.5	10	9	1	0.10
[30,34]	32	15	13	2	0.13
[35,39]	37	12	9	3	0.25
[40,44]	42	15	10	5	0.33
[45,49]	47	13	7	6	0.46
[50,54]	52	8	3	5	0.63
[55,59]	57	17	4	13	0.76
[60,69]	64.5	10	2	8	0.80

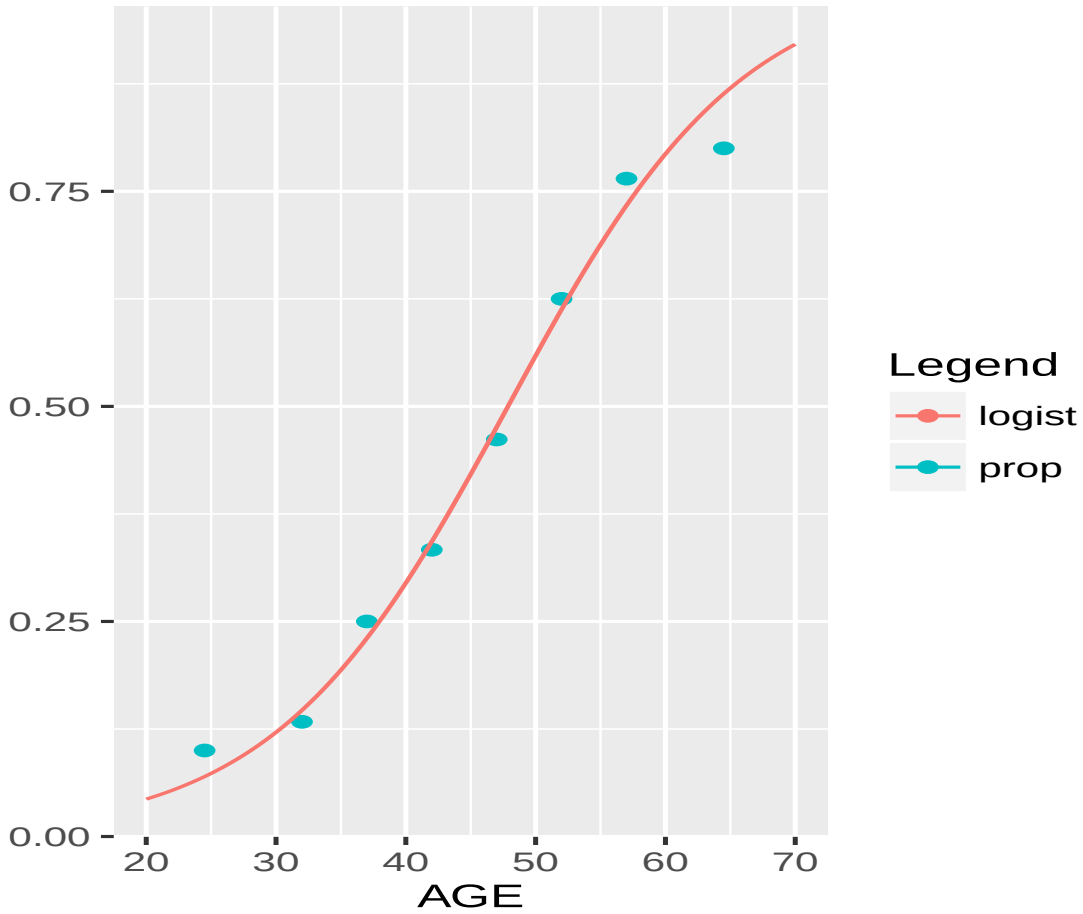

1.00
0.75
0.50
0.25
0.00
p
20
30
40
50
60
70
AGE
Legend
CHD
p (avec 8 part.)

Retour à l'exemple 2 : Extrait de sorties R

```
> CHD.logit = glm(CHD~AGE, family=binomial(link="logit"))
> summary(CHD.logit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

Modèle ajusté

$$\widehat{\pi}(x) = \frac{\exp(-5.30945 + 0.11092 \times age)}{1 + \exp(-5.30945 + 0.11092 \times age)}.$$

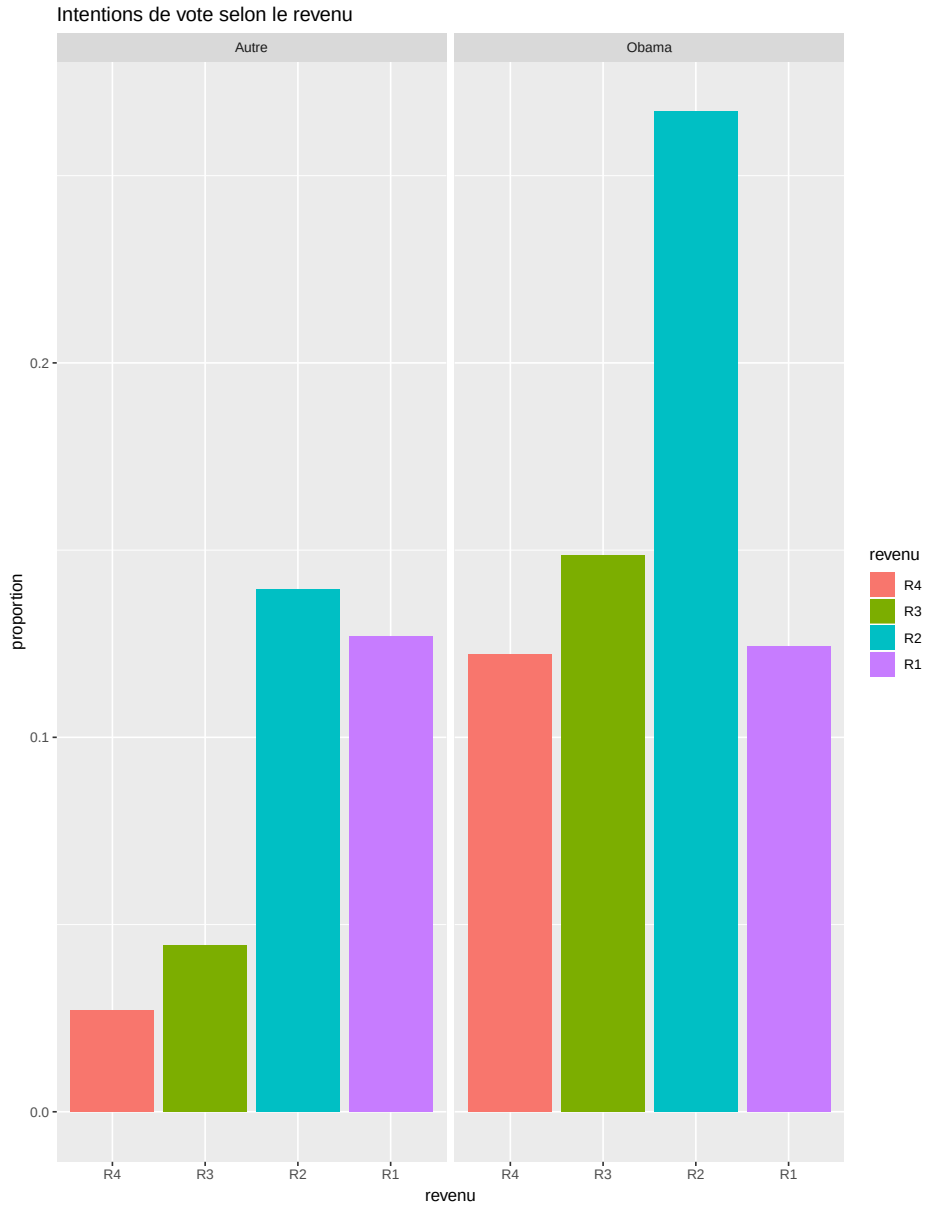


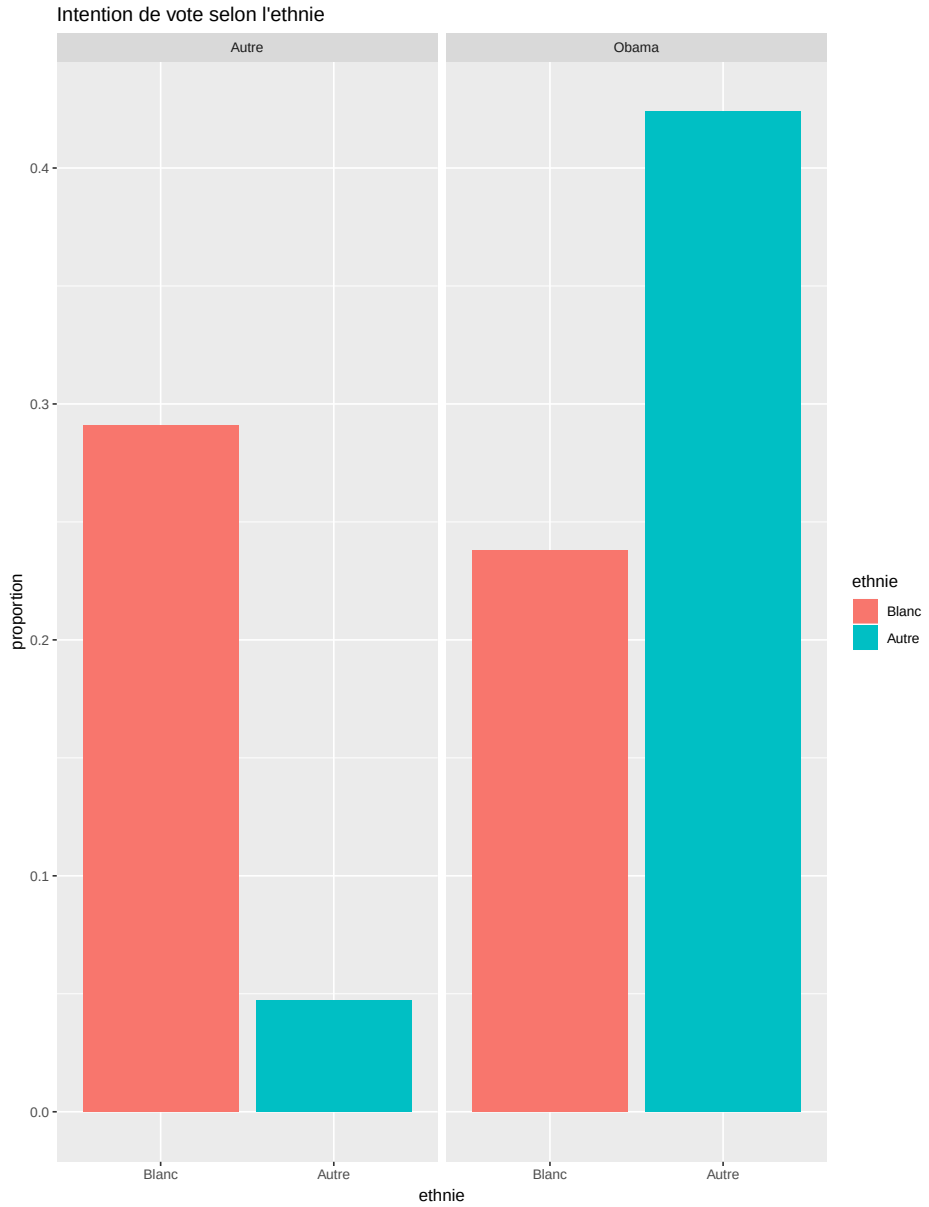
Exemple : Élection présidentielle américaine de 2008.

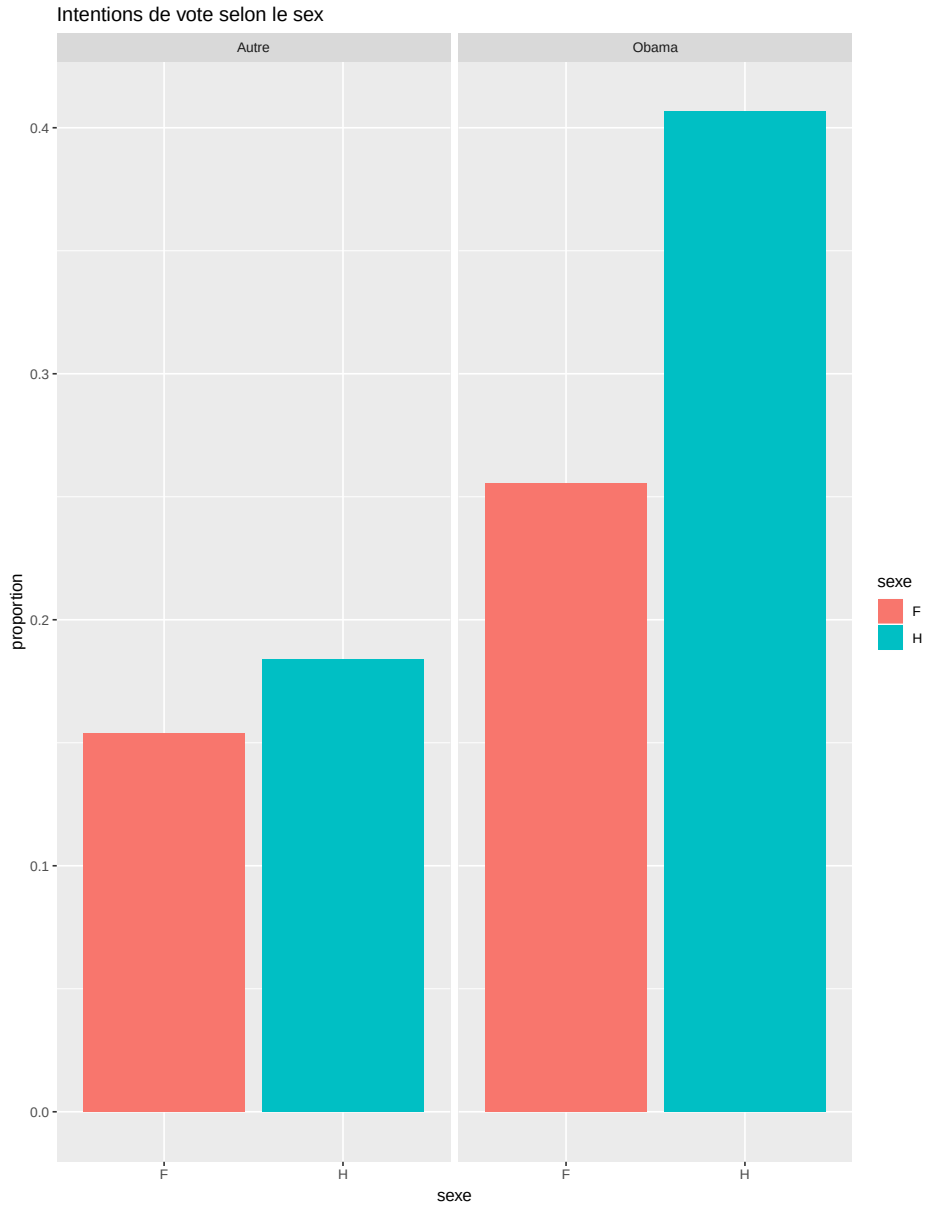
- On dispose d'un échantillon de taille 2322
- On s'intéresse pas exemple aux 4 variables
  - vote : 1 = vote pour Obama, 0 = ne vote pas pour Obama (Autre)
  - sexe : F= femmes , H = hommes
  - ethnie : Blanc et Autre (minorités ethniques)
  - revenue : R1 = très hauts revenus, R2 = hauts revenus, R3 = bas revenus, R4 = très bas revenus
- 5 électeurs au hasard

	sexe	ethnie	revenu	vote
446	H	Blanc	R3	Autre
59	H	Autre	R4	Obama
1258	F	Autre	R1	Autre
934	H	Blanc	R2	Autre
78	H	Autre	R2	Obama

```
sexe       ethnie     revenu          vote
F:590    Blanc:762    R4:215    Autre:487
H:851    Autre:679    R3:278    Obama:954
R2:586
R1:362
```


Intentions de vote selon le revenu
Autre
Obama
0.2
0.1
0.0
proportion
R4
R3
R2
R1
revenu
revenu
R4
R3
R2
R1


Intention de vote selon l'ethnie
Autre
Obama
0.4
0.3
0.2
0.1
0.0
proportion
Blanc
Autre
Blanc
Autre
ethnie
ethnie
Blanc
Autre


Intentions de vote selon le sex
Autre
Obama
0.4
0.3
0.2
0.1
0.0
proportion
F
H
F
H
sexe
sexe
F
H

- Données de l'élection présidentielle américaine de 2008.
- Est-ce que les électeurs de minorités ethniques (vs les Blancs), les gens moins favorisés économiquement (vs les plus favorisés) et les femmes (plus que les hommes) ont plus grande Pabilité d'appuyer le candidat démocrate Barack Obama ?.
- But : Estimer la Pabilité de voter pour Obama à partir des variables explicatives 'sexe', 'ethnie' et 'revenu'.

```
Call:  glm(formula = vote ~ sexe + ethnie + revenu,
family = binomial("logit"), data = ElectionUSA2008_A)

Coefficients:
(Intercept)         sexeH  ethnieAutre     revenuR3     revenuR2     revenuR1
0.3161            0.2036       2.2842      -0.2809      -0.5742      -1.0881

Degrees of Freedom: 1440 Total (i.e. Null);  1435 Residual
Null Deviance:     1844
Residual Deviance: 1455  AIC: 1467
```

- sexe="F",ethnie="Autre",revenu = "R4"
- sexe="H",ethnie="Autre",revenu = "R4"

```
1         2
0.9308788 0.9428834
```

- Une femme très défavorisée et appartenant à une minorité aurait voté pour Obama avec une Pa de 93,1%. La Pabilité pour un homme du même profil est de 94,3 %.

Cotes (odds) et rapports de cotes (odds ratios)

Dans le cas où la variable réponse Y est à valeurs dans $\{0, 1\}$ et $x = (x_1, x_2, \ldots, x_p)$, on définit :

La cote : $C(x) = \dfrac{\pi(x)}{1 - \pi(x)}$.

Le rapport de cotes : $OR := R(x', x) = \dfrac{C(x')}{C(x)}$.

Cotes (odds) et rapports de cotes (odds ratios)

Cas Simple : Supposons qu'on dispose d'une unique variable explicative X de type qualitative à deux modalités $\{0,1\}$.

Si l'on suppose que $\pi(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$ on a alors

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

avec β_0 et β_1 inconnus.

- La cote : $C(x) = \dfrac{\pi(x)}{1-\pi(x)}$.
- Coefficients estimés

$$\widehat{\beta_0} = log(C(0)) \quad \text{et} \quad \widehat{\beta_1} = \log(C(1)/C(0)) = \log(OR)$$

Rapports de cotes (odds ratios)

- Tableau de contingence

```
sex
vote       F   H
Autre    222 265
Obama    368 586
```

- Odds-ratio

```
(222*586)/(368*265)
[1] 1.334003
```

- L'odds-ratio >1 . Les hommes ont plus de chances de voter Obama que les femmes

Rapports de cotes (odds ratios)

- Expliquer le vote en fonction du sexe

```
Call:  glm(formula = vote ~ sexe,
family = binomial("logit"), data = ElectionUSA2008_A)

Coefficients:
(Intercept)         sexeH
0.5054         0.2882

Degrees of Freedom: 1440 Total (i.e. Null);  1439 Residu
Null Deviance:      1844
Residual Deviance: 1837  AIC: 1841
```

- Odds-ratio

```
exp(0.2882)
[1] 1.334024
```

Cotes (odds) et rapports de cotes (odds ratios)

- Le rapport de cotes : $OR := R(x', x) = \dfrac{C(x')}{C(x)}$.
- Si toutes les coordonnées de x et x' sont identique hormis la $j$ème, pour laquelle $x'_{[j]} = x_{[j]} + 1$, alors on a

$$R(x', x) = \exp(\beta_j)$$

- Pour una variable binaire $X^{(j)} \in \{0, 1\}$

$$\mathrm{OR}^{(j)} = \frac{\frac{\mathbb{P}(Y=1|X,X^{(j)}=1)}{\mathbb{P}(Y=0|X,X^{(j)}=1)}}{\frac{\mathbb{P}(Y=1|X,X^{(j)}=0)}{\mathbb{P}(Y=0|X,X^{(j)}=0)}} = \exp(\beta^{(j)})$$

- La définition est à peu près la même pour une variable réelle

Interpretation

- $\mathrm{OR}^{(j)} = 1$: $\mathbb{P}(Y = 1|X)$cne dépend pas de $X^{(j)}$.
- $\mathrm{OR}^{(j)} < 1$: $\mathbb{P}(Y = 1|X)$ diminue lorsque $X^{(j)}$ augmente.
- $\mathrm{OR}^{(j)} > 1$: $\mathbb{P}(Y = 1|X)$ augmente lorsque $X^{(j)}$ augmente

Références :

- An introduction to Generalized Linear Models, A.J. Dobson (2002)
- Statistiques avec $\mathbb{R}$, Pierre-André Cornillon et al. (2010), Presses universitaires de Rennes.
- Applied econometrics with R, Christian Kleiber et Achim Zeileis (2011), Springer.