Le modèle linéaire généralisé (logit, probit, ...)

Master 2 Recherche SES-IES Analyse de données

Ana Karina Fermin

Université Paris-Ouest-Nanterre-La Défense

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).

Exemples:

- Médecine : Y vaut 1 si le patient atteint la maladie, 0 sinon. La variable X est l'âge.
- Banque : Y vaut 1 si le client fait défaut sur sa dette. La variable X est par exemple l'âge, la profession, le montant moyen mensuel d'utilisation de la carte de crédit, le revenu du client,..., etc.
- Sociologie : Y vaut 1 si le fils est cadre, 0 sinon. La variable X est par exemple le niveau d'éducation du père.,

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. En pratique les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ doivent être déterminés à partir des données.

Modèle théorique

$$Y = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) + \varepsilon,$$

où le bruit ε est une variable aléatoire centrée.

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.

Dans ce cours on va juste utiliser et interpréter les résultats donnés par le logiciel R (vous n'avez pas besoin de connaitre les résultats théoriques de la log-vraissemblance associée au modèle) !!!

Notre objectif est modéliser

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

Modèle théorique

$$Y = \pi(X) + \varepsilon,$$

où $\pi(x) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$ et ε est centrée.

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 $t \in \mathbb{R}$,

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

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

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

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 : $OR = \dfrac{C(x')}{C(x)}$.

Cas de la régression logistique simple avec X qualitative

Cas Simple : Supposons qu'on dispose d'une unique variable explicative X de type qualitative à deux modalités {0,1}.
Nous avons fait un exemple à la main à l'aide d'un tableau de contingence pour les données de la mobilité sociale (voir vos notes de CM).

Si l'on suppose que

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

on a alors

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1$$

avec β_0 et β_1 inconnus.

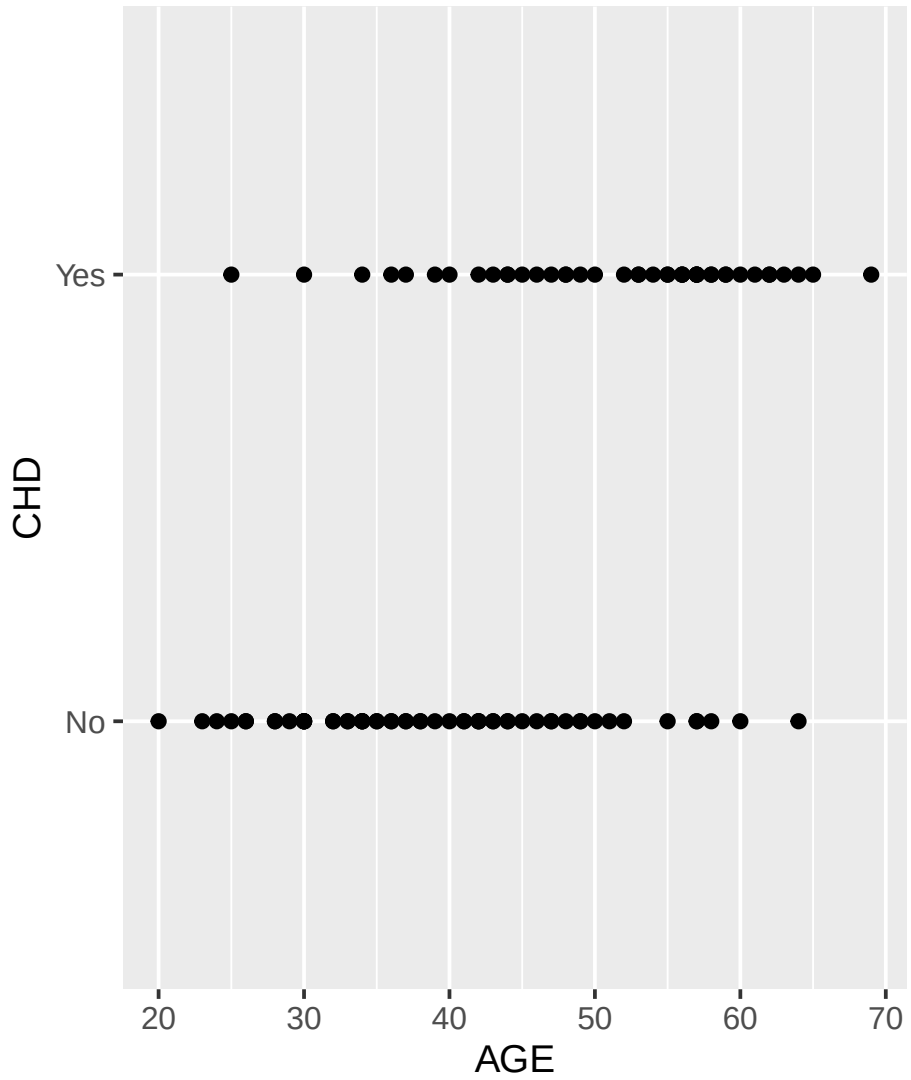
$$\widehat{\beta_0} = log(C(0)) \quad \text{et} \quad \widehat{\beta_1} = \log(C(1)/C(0)) = \log(OR)$$

Exemple 2 (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,

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

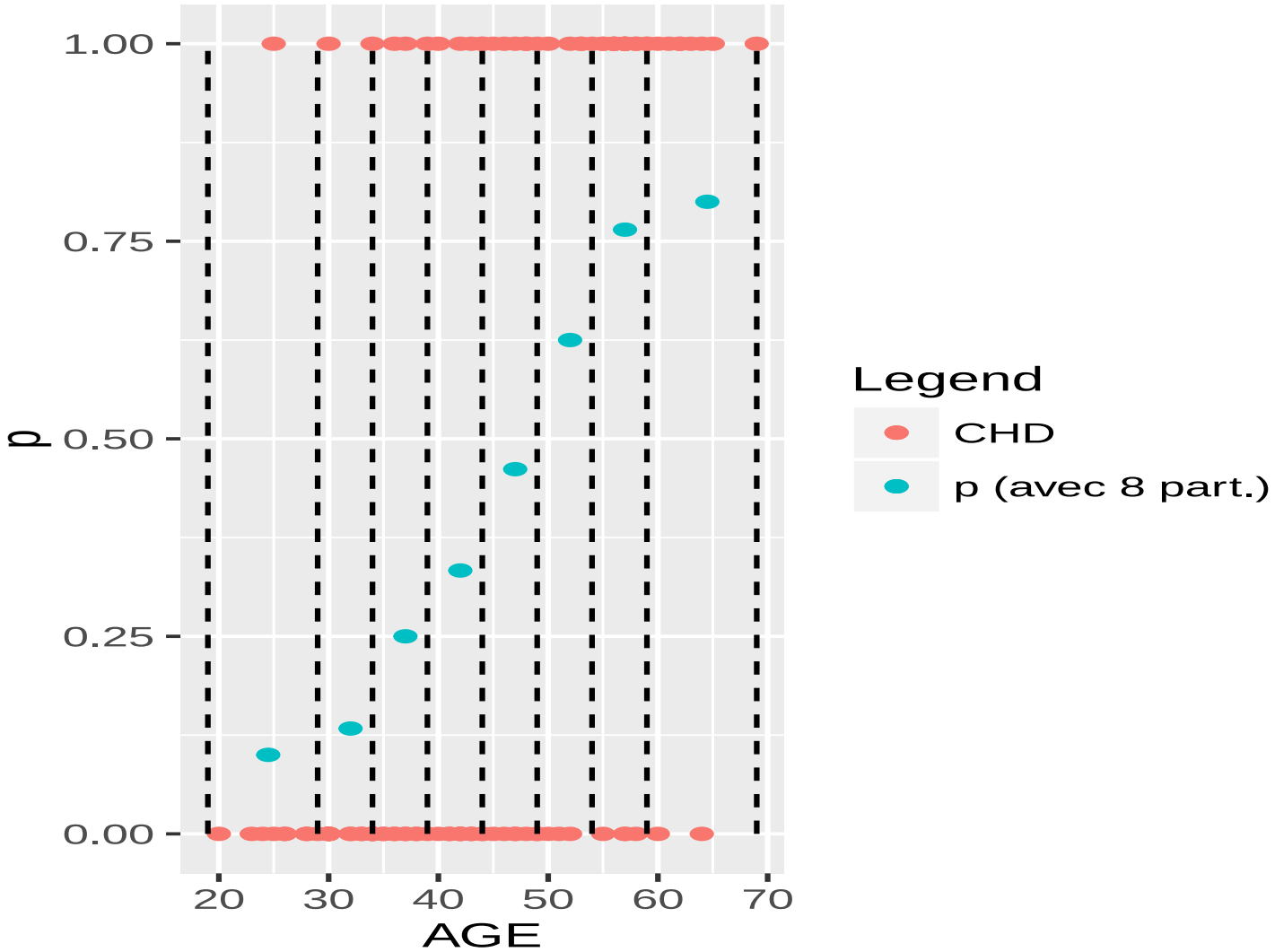
On dit dans ce cas que les données sont groupées. Sinon, on dit que les données sont individuelles

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.)

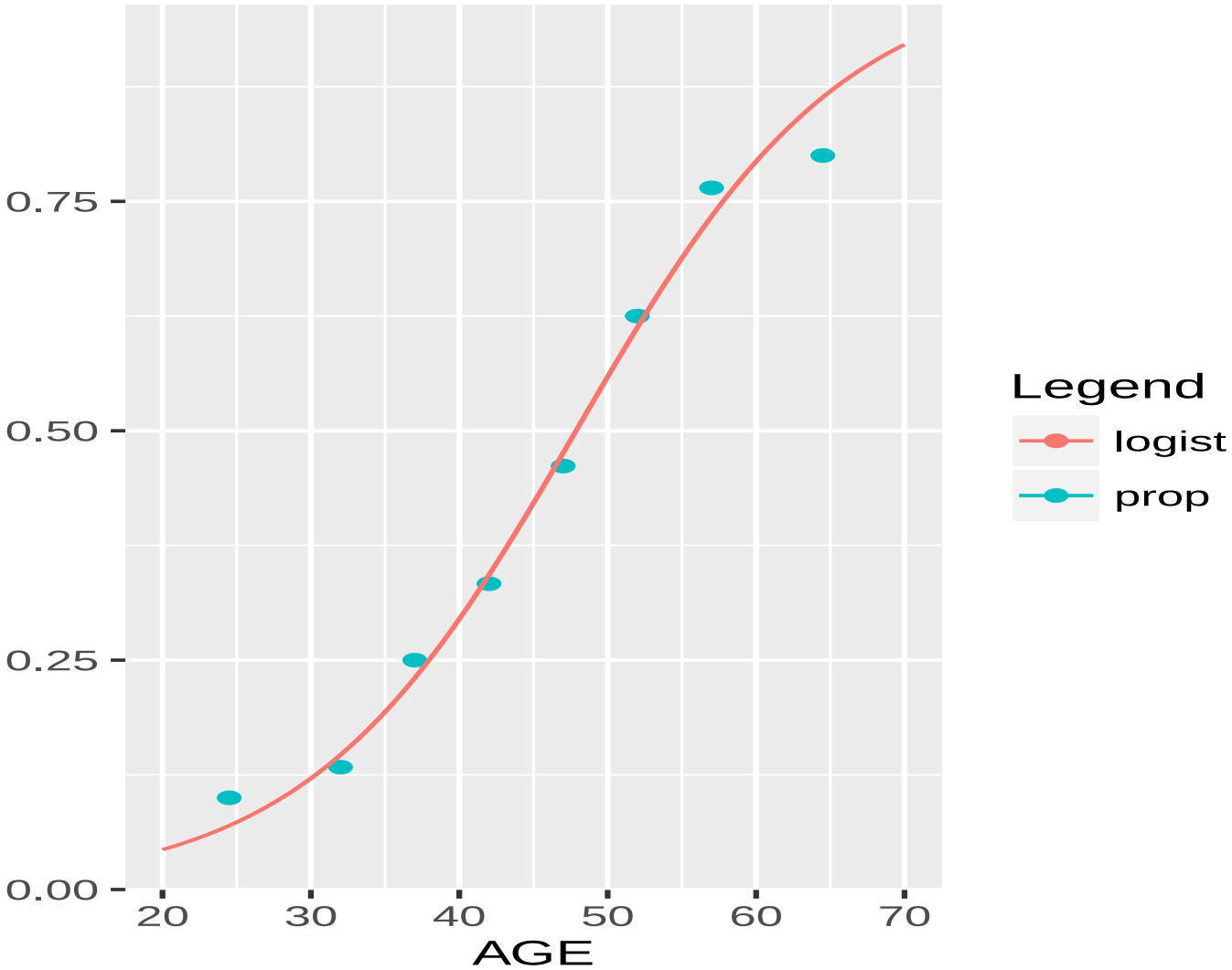
Retour à l'exemple 2 : Extrait de sorties R

```
> 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
```


0.75
0.50
0.25
0.00
20
30
40
50
60
70
AGE
Legend
logist
prop

Exemple 3 (cf. RIII)

Nous traitons un problème de défaut bancaire. Nous cherchons à déterminer quels clients seront en défaut sur leur dette de carte de crédit (ici `defaut` = 1 si le client fait défaut sur sa dette). La variable `defaut` est la variable réponse.

Nous disposons d'un échantillon de taille 10000 et 3 variables explicatives

- `student`: variable qualitative à 2 niveaux (student et non-student)
- `balance`: montant moyen mensuel d'utilisation de la carte de crédit
- `income`: revenu du client

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
student     -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.7  on 9997  degrees of freedom
AIC: 1577.7
```

Rappelons qu'on dispose d'un échantillon de taille $n = 10000$

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.