Améliorer l'usage des régressions logistiques en sociologie. Présentation d'un cas.

Muriel Epstein

Jean-Marc Bardet

Joseph Rynkiewicz

PLAN

- Contexte, sujet de l'étude et présentation de la base de données
- Rappels sur la régression logistique
- Premiers résultats avec des modèles classiques
- Le critère BIC
- Application à notre étude

CONTEXTE

Contexte sociologique

- Un collège qui ouvre en 2000 en Guyane, pas de modèle sur la réussite ou l'orientation scolaire
- Interrogation de l'ensemble des collégiens sur ce qu'est la réussite (451 réponses mais 432 réponses conservées dans l'étude présentée)

Présentation de la BDD

Pour réussir il faut

- Se débrouiller dans la vie
- Avoir de la chance
- Obtenir un travail
- Avoir du filon
- Obtenir des diplômes
- Apprendre le français
- Réussir à l'école

1 = tout à fait d'accord
2= d'accord
3= pas d'accord
4= pas du tout d'accord

Variables contextuelles

- Les parents ne s'intéressent pas à la scolarité des enfants
- Les profs sont toujours absents
- L'école est inutile
- Mes parents pensent que les diplômes sont nécessaires
- Mes parents peuvent m'aider pour mes devoirs

Variables autres

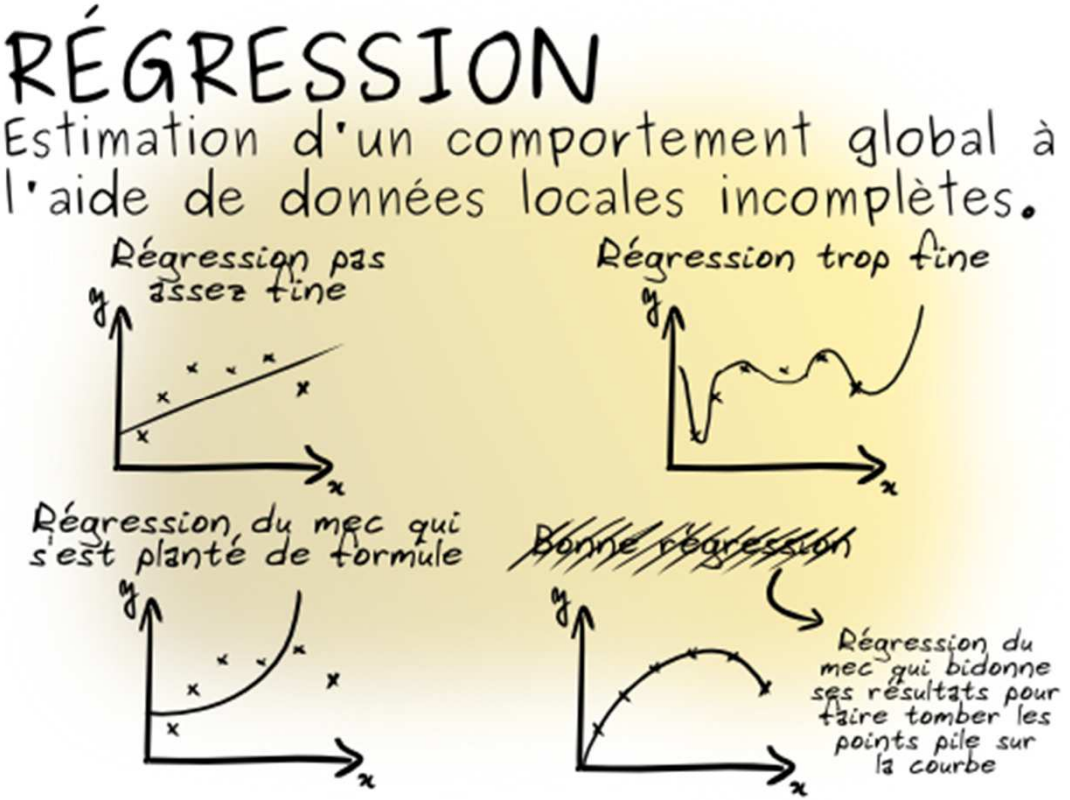
- Age
- Sexe
- Distance lieu résidence/collège
- Classe (spécifique ou générale)

Une différence de perception de la réussite filles et garçons

Tableau n°5 : ordre de priorité des éléments de réussite selon le sexe

Ordre de priorité	Ensemble N= 444	Filles N=232	Garçons=200
1	obtenir un ou des diplômes (1,28)	Avoir un emploi (1,34)	Fonder une famille (1,29)
2	Avoir un emploi (1,26)	Obtenir des diplômes (1,30)	Obtenir des diplômes (1,21)
3	Fonder une famille (1,05)	Avoir un logement à soi (0,89)	Avoir un emploi (1,19)
4	Avoir un logement à soi (0,85)	Avoir un emploi qui me plait (0,89)	Avoir un logement à soi (0,81)
5	Avoir un emploi qui me plait (0,82)	Fonder une famille (0,86)	Avoir un emploi qui me plait (0,71)
6	être riche (0,25)	Aller vivre dans une grande ville ou à l'étranger (0,31)	Pouvoir vivre dans sa ville d'origine (0,28)
7	Aller vivre dans une grande ville ou à l'étranger (0,25)	être riche (0,26)	être riche (0,26)
8	Pouvoir vivre dans sa ville d'origine (0,23)	Pouvoir vivre dans sa ville d'origine (0,19)	Aller vivre dans une grande ville ou à l'étranger (0,17)

[1] 19 valeurs manquantes à l'une ou l'autre question expliquent les 432 au lieu de 451



LA RÉGRESSION LOGISTIQUE

La régression logistique (1/3)

- On considère une variable à expliquer booléenne Y et des variables prédictives X. Avec I la variable indicatrice de Y=1

$$ln\frac{p(I|X)}{1-p(I|X)} = b_0+b_1X_1+\dots + b_jX_j$$

- Si on considère ω les individus de l'échantillon, la vraisemblance de l'échantillon (c'est-à-dire, pour des variables indépendantes, le produit de la probabilité d'appartenance des individus à un groupe) par

$$L = \prod_\omega P[Y(\omega) = 1|X(\omega)]^{Y(\omega)} \times [1 - P(Y(\omega)=1|X(\omega))]^{1-Y(\omega)}$$

Régression logistique (2/3)

Lorsqu'on maximise la vraisemblance, on trouve les coefficients β_j . On évalue classiquement la « réussite » d'une régression selon plusieurs moyens:

- La significativité du modèle (H_0 les b_i sont tous nuls contre au moins un des b_i ne l'est pas)
- La prédictivité du modèle (% de bonnes et mauvaises prédictions – courbe ROC)
- Le R^2 de Mac Fadden , forme de déviance (1- LL_M /LL_0) *[qui est une catastrophe]*

Régression logistique (3/3)

- AIC = -2LL + 2 x (J+1)
- BIC = -2LL + ln(n) x (J+1)
  - Avec -2LL la log vraisemblance
  - J+1 le nombre de paramètres à estimer
  - n le nombre d'individus

Comme pour les régressions linéaires, on considère qu'on travaille « toutes choses égales par ailleurs » et les colinéarités sont malvenues.

Il existe aussi des indicateurs variable par variable (test de Wald)

PREMIERS RÉSULTATS

Modèles pour expliquer « Réussir c'est fonder une famille »

		Modèle 1	Modèle 2	Modèle 3	Modèle 4	
Constante		ns	0,61**	0,53**	ns	0,13***
Sexe	Fille	0,490 ***	0,43***	0,40***		
Ref garçons					Garçon	Fille
Retard	1 an		1,77 *	1,74*	ns	ns
Ref à l'heure	2 ans		2,77 ***	2,34**	ns	3,76 **
	3 ans et plus		1,98 **	ns	ns	ns
Age	Moins de 13 ans			ns	ns	ns
Ref 14 ans	15 ans			2,14**	ns	2,77**
	Plus de 16 ans			4,21***	ns	7,58***
N		424	424	424	196	228
R^2 de Nagelkerke		4%	7%	10%	3%	10%
Prédictivité		59,2%	61,6%	62,0%	55,6%	68,0%

Tableau n°7 : Odds ratio de la régression logistique binaire sur « Réussir c'est fonder une famille »
*** risque d'erreur inférieur à 1% ** risque d'erreur inférieur à 5% * risque d'erreur inférieur à 10%
ns résultat non significatif

Premières conclusions sociologiques

- Fonder une famille est plus important pour les garçons que pour les filles
- Pour les garçons, l'importance accordée au fait de fonder une famille ne dépend d'aucun facteur scolaire présents en base de données
- Pour les filles, plus elles sont en échec scolaire et plus elles sont âgées par rapport à la norme, plus il y a de chances qu'elles veulent fonder une famille (*ce qui inverse le lien de causalité souvent donné entre l'abandon d'étude qui serait une conséquence de la maternité précoce*)

Premières conclusions mathématiques

- Un usage des régressions logistiques limité en sociologie qui reste largement fondé sur le R^2
- 432 observations et 30 variables conservées juste pour l'explication (dont la distance au collège, les différentes perceptions de la réussite)
- Quel serait le meilleur modèle, comment le chercher et pourquoi?

LE CRITÈRE BIC

Le critère BIC est plus pertinent

- Articles de Schwarz 1978 Raftery 1995 sur l'usage du BIC plutôt que du R^2

- Le BIC permet de choisir asymptotiquement le modèle le plus probable

APPLICATION

Premier modèle

(GLMulti Avec toutes les variables)

ReussirpourvousFonderunefamille ~ 1+Avoirdelachance + Obteniruntravail + Parentsetscolarite + Inutilite + Necessitedesdiplomes + Aideauxdevoirs +Sexe

Deviance Residuals:

Min 1Q Median 3Q Max -2.2524 -0.5607 -0.1104 0.5939 3.2212

Coefficients:	Estimate	value Pr(>|z|)
(Intercept)	6.0041965	2.62e-09 ***
Avoirdelachance2	-1.5618596	8.78e-05 ***
Avoirdelachance3	-2.9048304	1.74e-07 ***
Avoirdelachance4	-2.5489539	0.000214 ***
Obteniruntravail2	1.4840142	3.848 0.000119 ***
Obteniruntravail3	-3.8139496	0.003963 **
Obteniruntravail4	0.9783052	0.026566 *
Parentsetscolarite2	-3.7476286	8.56e-05 ***
Parentsetscolarite3	-0.6070015	0.395542
Parentsetscolarite4	-3.2139488	2.90e-06 ***
Inutilite2	-0.0007603	0.999337
Inutilite3	-3.1594171	0.000460 ***
Inutilite4	-4.5180253	1.42e-08 ***
Necessitedesdiplomes2	0.1553621	0.668993
Necessitedesdiplomes3	-2.7667646	0.000447 ***
Necessitedesdiplomes4	-0.2479703	0.681300
Aideauxdevoirs2	-0.3359720	0.551419
Aideauxdevoirs3	4.2830120	3.48e-10 ***
Aideauxdevoirs4	3.0760767	1.47e-06 ***
Sexe1	-1.5570020	3.50e-07 ***

Variable	Interprétation
Avoir de la chance	Penser que pour réussir il faut avoir de la chance est en opposition avec la vision que réussir c'est fonder une famille
Obtenir un travail	Pas d'interprétation logique
Parents et scolarité	Plus les parents s'intéressent à la scolarité, moins réussir c'est fonder une famille
Inutilité	Ceux qui pensent que l'école n'est pas inutile veulent moins fonder une famille
Nécessité des diplômes	Pas d'interprétation logique
Aides aux devoirs	Plus les parents peuvent aider les enfants à faire les devoirs, plus réussir c'est fonder une famille
Sexe	Les garçons veulent plus fonder une famille que les filles

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1)

Null deviance: 588.13 on 431 degrees of freedom Residual deviance: 326.09 on 412 degrees of freedom AIC: 366.09 Number of Fisher Scoring iterations: 6

- A l’issue desquels on n’obtient des modèles peu interprétables sociologiquement d’où recodage:
  - Distance
  - Classe
  - Nature des données

Second Modèle

(GM multi avec les variables recodées)

ReussirpourvousFonderunefamille ~ Avoirdelachance + Obteniruntravail + Parentsetscolarite + Inutilite + Aideauxdevoirs + Necessitedesdiplomes +Sexe+Age+classe2

```
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.3915  -0.5363  -0.1071   0.5242   3.1850

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.47020    3.29856  -1.658 0.097246 .
Avoirdelachance2   -1.75007    0.42510  -4.117 3.84e-05 ***
Avoirdelachance3   -3.07958    0.59114  -5.210 1.89e-07 ***
Avoirdelachance4   -2.86145    0.76303  -3.750 0.000177 ***
Obteniruntravail2   1.59204    0.40381   3.943 8.06e-05 ***
Obteniruntravail3  -3.93549    1.26376  -3.114 0.001845 **
Obteniruntravail4   0.95362    0.49116   1.942 0.052192 .
Parentsetscolarite2  -4.42966    1.06291  -4.167 3.08e-05 ***
Parentsetscolarite3  -1.05317    0.76145  -1.383 0.166633
Parentsetscolarite4  -3.44247    0.73144  -4.706 2.52e-06 ***
Inutilite2          0.29008    0.96220   0.301 0.763047
Inutilite3         -3.98234    0.95329  -4.177 2.95e-05 ***
Inutilite4         -4.51597    0.79852  -5.655 1.55e-08 ***
Necessitedesdiplomes2 0.12225    0.37402   0.327 0.743775
Necessitedesdiplomes3 -4.21796    1.10592  -3.814 0.000137 ***
Necessitedesdiplomes4 -0.36340    0.64594  -0.563 0.573708
Aideauxdevoirs2     0.04899    0.60072   0.082 0.935002
Aideauxdevoirs3     4.48357    0.73814   6.074 1.25e-09 ***
Aideauxdevoirs4     3.09543    0.67820   4.564 5.01e-06 ***
Sexe1              -1.64814    0.32790  -5.026 5.00e-07 ***
Age                0.82332    0.22223   3.705 0.000212 ***
classe14           0.13792    0.48333   0.285 0.775369
classe15          -0.67537    0.49770  -1.357 0.174787
classe110          3.66712    1.34287   2.731 0.006318 **
classe16           1.93090    0.80497   2.399 0.016453 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 588.13  on 431  degrees of freedom
Residual deviance: 302.56  on 407  degrees of freedom
AIC: 352.56
Number of Fisher Scoring iterations: 6
```

Variable	Interprétation
Avoir de la chance	Penser que pour réussir il faut avoir de la chance est en opposition avec la vision que réussir c'est fonder une famille
Obtenir un travail	Pas d'interprétation logique
Parents et scolarité	Plus les parents s'intéressent à la scolarité, moins réussir c'est fonder une famille
Inutilité	Ceux qui pensent que l'école n'est pas inutile veulent moins fonder une famille
Nécessité des diplômes	Pas d'interprétation logique
Aides aux devoirs	Plus les parents peuvent aider les enfants à faire les devoirs, plus réussir c'est fonder une famille
Sexe	Les garçons veulent plus fonder une famille que les filles
Age	Envie de fonder une famille augmente avec l'âge
Classe	Non significatif

Conclusion

- Connaitre ses données reste indispensable
- Le recodage est efficace
- Concordance des méthodes sur les « grands » résultats
- Apparition de quelques variables qui ont du sens sociologiquement (l'intérêt des parents pour l'école est sûrement un proxi de la PCS)