

RÉGRESSIONS LOGISTIQUES

MASTER 1

PLAN DU COURS

Régression logistique binaire simple (chd)

Régression logistique binaire multiple

- données individuelles (faillite, bébé)
- données agrégées (job satisfaction)

Régression logistique ordinaire (bordeaux)

- pentes égales
- partiellement à pentes égales

Régression logistique multinomiale (bordeaux, alligator)

2

A. LA RÉGRESSION LOGISTIQUE BINAIRE

Les données

Y = variable à expliquer binaire

X_1, \dots, X_k = variables explicatives numériques
ou binaires (indicatrices de modalités)

- Régression logistique binaire simple ($k = 1$)
- Régression logistique binaire multiple ($k > 1$)

3

I. LA RÉGRESSION LOGISTIQUE BINAIRE SIMPLE

Variable dépendante : $Y = 0$ ou 1

Variable indépendante : X quantitative ou binaire

Objectif : Modéliser

$$\pi(x) = \text{Prob}(Y = 1/X = x)$$

- Le modèle linéaire $\pi(x) = \beta_0 + \beta_1 x$ convient mal lorsque X est continue.
- Le modèle logistique est plus naturel.

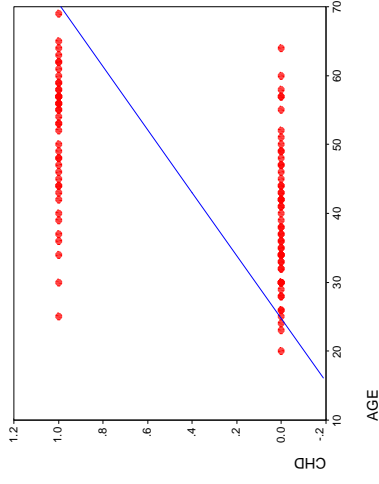
4

Exemple : Age and Coronary Heart Disease Status (CHD)

Les données

ID	AGRP	AGE	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	0
...
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

Plot of CHD by Age

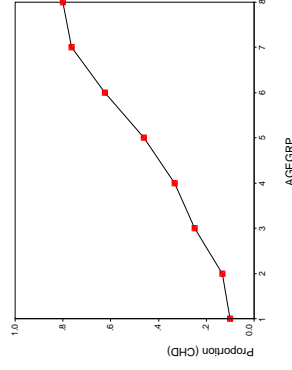


DESCRIPTION DES DONNÉES REGROUPÉES PAR CLASSE D'AGE

Tableau des effectifs de CHD par classe d'âge

Age Group	n	CHD absent	CHD present	Mean (Proportion)
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43

Graphique des proportions de CHD par classe d'âge



LE MODÈLE LOGISTIQUE

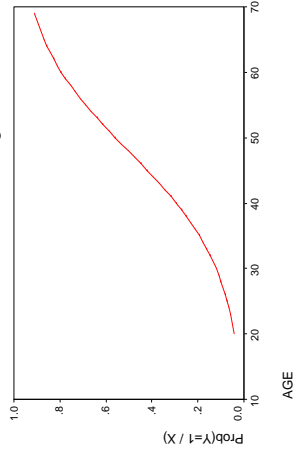
$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

ou

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

Fonction de lien : Logit

Probabilité d'une maladie cardiaque en fonction de l'âge



FONCTIONS DE LIEN

- Fonction logit

$$g(p) = \log(p / (1 - p))$$
 - Fonction normit ou probit

$$g(p) = \Phi^{-1}(p)$$
- où Φ est la fonction de répartition de la loi normale réduite
- Fonction « complementary log-log »

$$g(p) = \log(-\log(1-p))$$

6

ESTIMATION DES PARAMÈTRES DU MODÈLE LOGISTIQUE

Les données

\mathbf{X}	\mathbf{Y}
\mathbf{x}_1	y_1
\vdots	\vdots
\mathbf{x}_i	y_i
\vdots	\vdots
\mathbf{x}_n	y_n

$y_i = 1$ si caractère présent,
 0 sinon

Le modèle

$$\pi(x_i) = P(Y = 1 / X = x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

11

VRAISEMBLANCE DES DONNÉES

Probabilité d'observer les données

$[(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)]$

$$= \prod_{i=1}^n \text{Prob}(Y = y_i / X = x_i)$$

$$= \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

$$= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i}$$

$$= \ell(\beta_0, \beta_1)$$

11

LOG-VRAISEMBLANCE

$$L(\beta_0, \beta_1) = \text{Log}(\ell(\beta_0, \beta_1)) = \text{Log} \left[\prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right]$$

$$= \sum_{i=1}^n y_i \text{Log} \left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right) + \text{Log}(1 - \pi(x_i))$$

$$= \sum_{i=1}^n y_i (\beta_0 + \beta_1 x_i) - \text{Log}(1 + \exp(\beta_0 + \beta_1 x_i))$$

12

ESTIMATION DU MAXIMUM DE VRAISEMBLANCE

On cherche $\hat{\beta}_0$ et $\hat{\beta}_1$ maximisant la Log-vraisemblance $L(\hat{\beta}_0, \hat{\beta}_1)$.

$$\text{La matrice } V(\hat{\beta}) = \begin{bmatrix} V(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_1) \end{bmatrix}$$

est estimée par la matrice $\left[-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} \right]_{\beta=\hat{\beta}}$

13

RÉSULTATS

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.353	.254	.341

Variables in the Equation

Step	AGE	B	S.E.	Wald	df	Sig.	Exp(B)
1	Constant	-5.309	1.134	21.935	1	.000	.005
	AGE	.111	.024	21.254	1	.000	1.117

a. Variable(s) entered on step 1: AGE.

Test LRT pour $H_0 : \beta_1 = 0$

Omnibus Tests of Model Coefficients

Step 1	Model	Chi-square	df	Sig.
		29.310	1	.000

14

RÉSULTATS

Variable	Intercept	age
Intercept	1.285173	-0.02668
age	-0.02668	0.000579

Ecart-type de la constante = 1.285173^{1/2} = 1.134

Ecart-type de la pente = .000579^{1/2} = .024

Covariance entre la constante et la pente = -.02668

15

TEST DE WALD

Le modèle

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Test

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned}$$

Statistique utilisée

$$\text{Wald} = \frac{\hat{\beta}_1^2}{s_1^2}$$

Décision de rejeter H_0 au risque α

Rejet de H_0 si $\text{Wald} \geq \chi_{1-\alpha}^2(1)$

ou $\text{NS} = P(\chi^2(1) \geq \text{Wald}) \leq \alpha$

91

TEST LRT

Le modèle

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned}$$

Test

Statistique utilisée

$$\Lambda = [-2L(Cste)] - [-2L(Cste, X)]$$

Décision de rejeter H_0 au risque α

$$\text{Rejet de } H_0 \text{ si } \Lambda \geq \chi^2_{1-\alpha}(1) \quad \text{ou} \quad \text{NS} = P(\chi^2(1) \geq \Lambda) \leq \alpha$$

INTERVALLE DE CONFIANCE DE $\pi(x)$ AU NIVEAU 95%

De

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \text{Var}(\hat{\beta}_0) + x^2 \text{Var}(\hat{\beta}_1) + 2x \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

on déduit l'intervalle de confiance de

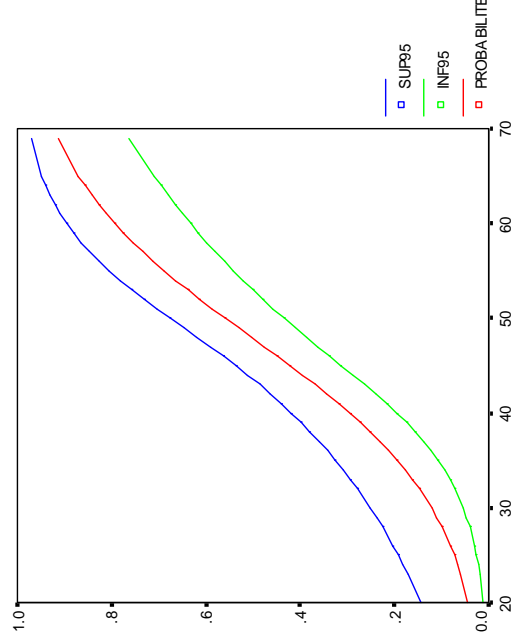
$$\left[\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x - 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x - 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}, \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}} \right]$$

Case Summaries*

	AGE	PROBABILITE CALCULE	INFR95	SUPR95
1	20.0	.04	.01	.14
2	23.0	.06	.02	.17
3	24.0	.07	.02	.18
4	25.0	.07	.03	.19
5	25.0	.07	.03	.19
6	26.0	.08	.03	.20
7	26.0	.08	.03	.20
8	28.0	.10	.04	.23
9	28.0	.10	.04	.23
10	29.0	.11	.05	.24
11	30.0	.12	.05	.25
12	30.0	.12	.05	.25
13	30.0	.12	.05	.25
14	30.0	.12	.05	.25
15	30.0	.12	.05	.25
16	30.0	.12	.05	.25
17	32.0	.15	.07	.28
18	32.0	.15	.07	.28
19	33.0	.16	.08	.29
20	33.0	.16	.08	.29
21	34.0	.18	.09	.31
22	34.0	.18	.09	.31
23	34.0	.18	.09	.31
24	34.0	.18	.09	.31
25	34.0	.18	.09	.31
26	34.0	.18	.09	.31
27	35.0	.19	.11	.33
28	35.0	.19	.11	.33
29	36.0	.21	.12	.34
30	36.0	.21	.12	.34
Total	N	30	30	30

a. Limited to first 30 cases.

INTERVALLE DE CONFIANCE DE $\pi(x)$ AU NIVEAU 95%



Comparaison entre les proportions observées et théoriques

Proportion observée :

$$\sum_{i \in \text{Classe}} y_i / n_{\text{Classe}}$$

Proportion théorique :

$$\sum_{i \in \text{Classe}} \hat{\pi}_i / n_{\text{Classe}}$$

puisque $E(y_i) = \pi_i$
estimé par $\hat{\pi}_i$

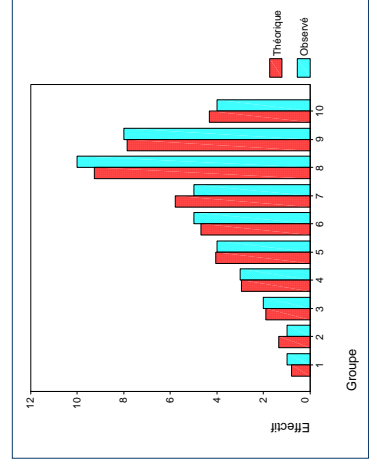
Report

AGEGRP	Mean	Maladie cardiaque	Predicted probability
1	Mean N	.1000 10	.0787086 10
2	Mean N	.1333 15	.1484562 15
3	Mean N	.2500 12	.2299070 12
4	Mean N	.3333 15	.3519639 15
5	Mean N	.4615 13	.4824845 13
6	Mean N	.6250 8	.6087623 8
7	Mean N	.7647 17	.7302152 17
8	Mean N	.8000 10	.8391673 10
Total	Mean N	.4300 100	.4300000 100

21

TEST DE HOSMER & LEMESHOW (GOODNESS OF FIT TEST)

Les données sont rangées par ordre croissant des probabilités calculées à l'aide du modèle, puis partagées en 10 groupes au plus. **Ce test est malheureusement peu puissant.**

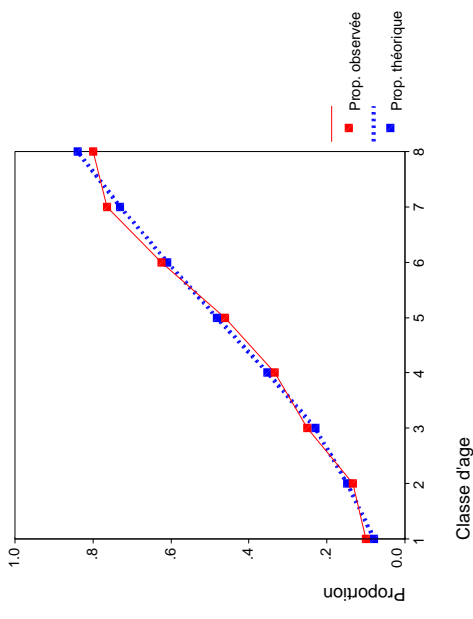


Le test du khi-deux est utilisé pour comparer les effectifs observés ($\sum_{i \in \text{Classe}} y_i$) aux effectifs théoriques ($\sum_{i \in \text{Classe}} \hat{\pi}_i$).

Nb de degrés de liberté = Nb de groupes - 2

23

Comparaison entre les proportions observées et théoriques



22

TEST DE HOSMER & LEMESHOW

Contingency Table for Hosmer and Lemeshow Test

Step	Maladie cardiaque = chd=no		Maladie cardiaque = chd=yes		Total
	Observed	Expected	Observed	Expected	
1	9	9.213	1	.787	10
2	9	8.657	1	1.343	10
3	8	8.095	2	1.905	10
4	8	8.037	3	2.963	11
5	7	6.947	4	4.053	11
6	5	5.322	5	4.678	10
7	5	4.200	5	5.800	10
8	3	3.736	10	9.264	13
9	2	2.134	8	7.866	10
10	1	.661	4	4.339	5

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.890	8	.999

24

MESURE DE LA QUALITÉ DE LA MODÉLISATION

R² de Cox & Snell

$$R^2 = 1 - \left[\frac{\ell(cte)}{\ell(cte, X)} \right]^{\frac{2}{n}}$$

Pseudo R² (McFadden)

$$Pseudo - R^2 = 1 - \left[\frac{-2L(cte, X)}{-2L(cte)} \right]$$

$$Max R^2 = 1 - \left[\ell(cte) \right]^{\frac{2}{n}}$$

R² ajusté de Nagelkerke

$$R^2_{adj} = \frac{R^2}{R^2_{max}}$$

25

TABLEAU DE CLASSIFICATION

Une observation *i* est affectée à la classe [Y=1] si

$$\hat{\pi}_i \geq c.$$

Tableau de classification (**c = 0.5**)

TABLE OF CHD BY PREDICTS			
CHD	PREDICTS		Total
	0	1	
Frequency	0	1	Total
0	45	12	57
1	14	29	43
Total	59	41	100

Sensibilité = 29/43
Spécificité = 45/57

taux de faux positifs = 12/41
taux de faux négatifs = 14/59

26

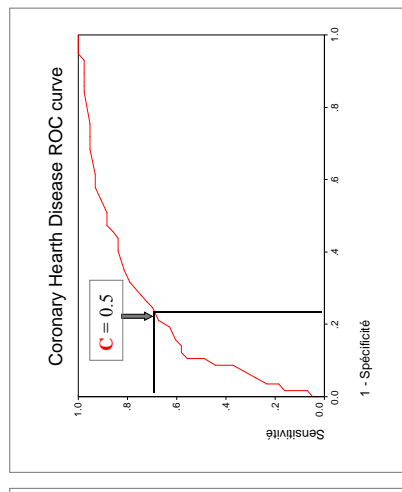
GRAPHIQUE ROC (RECEIVER OPERATING CHARACTERISTIC)

Sensibilité : capacité à prédire un événement
Spécificité : capacité à prédire un non-événement

Graphique ROC :

$$y = \text{Sensibilité}(c)$$

$$x = 1 - \text{Spécificité}$$



(c)

L'aire sous la courbe ROC est une mesure du pouvoir prédictif de la variable X. Ici cette surface est égale à 0.8.

28

OBJECTIFS

Sensibilité = capacité à diagnostiquer les malades parmi les malades

Spécificité = capacité à reconnaître les non-malades parmi les non-malades

1 - Spécificité = risque de diagnostiquer un malade chez les non-malades.



Trouver un compromis acceptable entre forte sensibilité et forte spécificité.

27

COEFFICIENTS D'ASSOCIATION ENTRE LES PROBABILITÉS CALCULÉES ET LES RÉPONSES OBSERVÉES

N = effectif total

t = nombre de paires avec
des réponses différentes

$$= nb(0) * nb(1)$$

nc = nombre de paires
concordantes ($y_i < y_j$ et

$$\hat{\pi}_i < \hat{\pi}_j$$

nd = nombre de paires
discordantes ($y_i < y_j$ et

$$\hat{\pi}_i > \hat{\pi}_j$$

t - nc - nd = Nb d'ex-aequo

$$(y_i < y_j \text{ et } \hat{\pi}_i = \hat{\pi}_j)$$

$$D \text{ de Somer} = (nc - nd) / t$$

$$\text{Gamma} = (nc - nd) / (nc + nd)$$

$$\text{Tau-a} = (nc - nd) / (.5N(N-1))$$

$$c = (nc + .5(t - nc - nd)) / t$$

c = aire sous la courbe

ROC

29

Analyse des résidus données individuelles

Résidu de Pearson (Standardized Residual)

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

à comparer à 2 en valeur absolue

30

Autres statistiques pour l'analyse des résidus

Déviance :

$$D = -2 \log \ell = \sum d_i^2$$

Résidu déviance (Deviance)

$$d_i = \text{signe}(y_i - \hat{\pi}_i) \sqrt{-2 \log(\text{Prob}_{\text{estimée}}[Y = y_i / X = x_i])}$$

à comparer à 2 en valeur absolue

Influence de chaque observation sur la déviance (DiffDev)

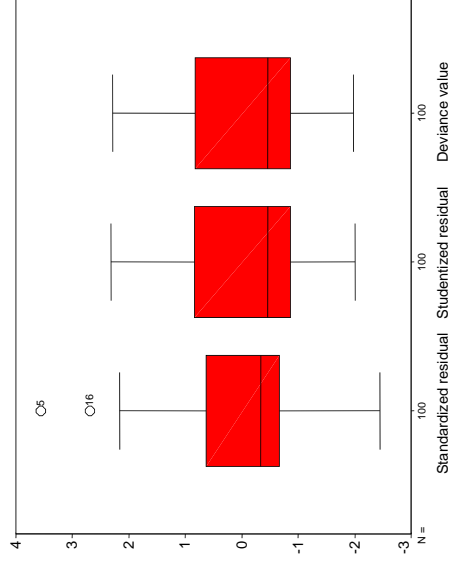
$$\Delta_i D = D(\text{toutes les obs.}) - D(\text{toutes les obs. sauf l'obs. } i)$$

Studentized residual :

$$\text{signe}(y_i - \hat{\pi}_i) \sqrt{\Delta D_i}$$

31

ANALYSE DES RÉSIDUS



32

II. LA RÉGRESSION LOGISTIQUE MULTIPLE

EXEMPLE : PRÉVISION DE FAILLITE

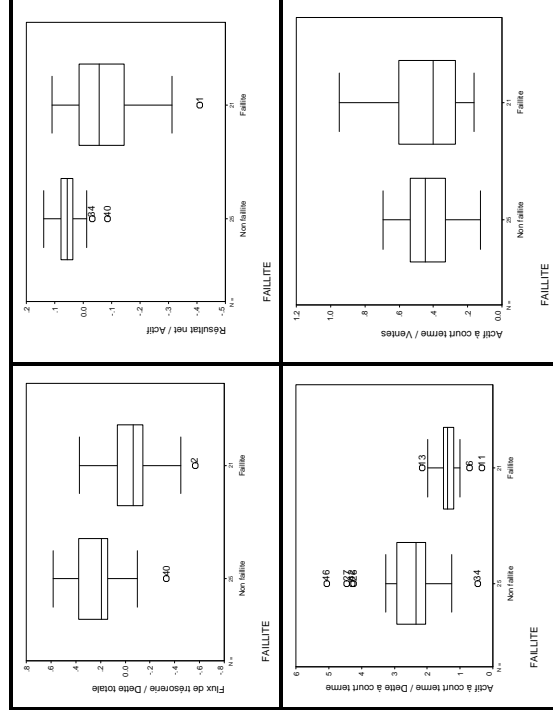
Les données

Les ratios suivants sont observés sur 46 entreprises :

- X_1 = Flux de trésorerie / Dette totale
- X_2 = Resultat net / Actif
- X_3 = Actif à court terme / Dette à court terme
- X_4 = Actif à court terme / Ventas
- Y = F si faillite, NF sinon

Deux ans après 21 de ces entreprises ont fait faillite et 25 sont restées en bonne santé financière.

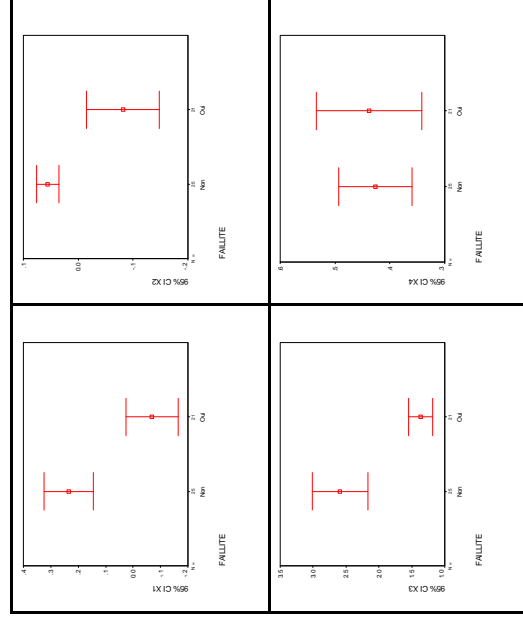
Boîtes à moustaches des ratios financiers selon le critère de Faillite



	cash flow / total debt	net income / total assets	current assets / current liabilities	current assets / net sales	FAILLITE
1	-0,6	-4,1	1,09	-4,5	F
2	-0,6	0,5	1,01	-0,5	F
3	-0,6	0,2	1,01	-0,4	F
4	-0,7	-0,9	1,45	-2,6	F
5	-1,0	-0,9	1,56	-0,7	F
6	-1,4	-0,7	1,71	-2,8	F
7	0,4	0,1	1,50	-7,1	F
8	0,4	0,1	1,50	-7,1	F
9	0,7	-0,1	1,37	-3,4	F
10	-1,4	-1,4	1,42	-4,3	F
11	-2,3	-3,0	1,18	-1,8	F
12	0,7	0,2	1,31	-2,5	F
13	0,7	0,2	1,31	-2,5	F
14	0,8	0,0	2,15	-7,0	F
15	-3,8	-2,3	1,44	0,7	F
16	1,5	0,5	1,88	2,7	F
17	3,7	-1,1	1,99	-3,6	F
18	0,6	-0,8	1,51	-4,2	F
19	0,6	0,3	1,69	-0,5	F
20	0,6	0,3	1,69	-0,5	F
21	0,6	0,3	1,69	-0,5	F
22	-2,8	-2,7	1,27	5,1	F
23	0,8	0,2	2,01	-5,4	NF
24	3,8	1,1	3,27	-5,3	NF
25	3,8	1,1	3,27	-5,3	NF
26	3,8	1,1	3,27	-5,3	NF
27	3,1	0,5	4,24	6,9	NF
28	3,1	0,5	4,45	6,9	NF
29	-0,2	0,2	2,52	6,9	NF
30	-0,2	0,2	2,52	6,9	NF
31	2,3	0,3	2,35	-4,0	NF
32	1,5	0,5	2,17	-5,5	NF
33	-1,0	-0,1	2,50	-5,8	NF
34	-1,4	-0,3	4,6	-2,6	NF
35	1,4	0,7	2,61	-5,2	NF
36	1,4	0,7	2,61	-5,2	NF
37	1,4	0,7	2,61	-5,2	NF
38	2,0	0,6	1,84	3,9	NF
39	5,4	-1,1	2,33	-4,8	NF
40	-3,3	-0,9	3,01	-4,7	NF
41	-4,8	-0,9	1,24	-1,8	NF
42	3,0	0,1	1,42	-4,4	NF
43	3,0	0,1	1,42	-4,4	NF
44	4,7	-1,4	2,92	-4,5	NF
45	1,7	-0,4	2,45	-1,4	NF
46	0,5	0,4	5,05	-1,3	NF

a. Limited to first 100 cases.

Intervalle de confiance des moyennes des ratios financiers selon le critère de Faillite



RÉGRESSIONS LOGISTIQUES SIMPLES DE Y SUR LES RATIOS X

Variable	Coefficient β_i	WALD	NS	R^2 de Nagelkerke
X ₁	-7.526	9.824	.002	.466
X ₂	-19.493	8.539	.003	.466
X ₃	-3.382	11.75	.001	.611
X ₄	.354	.040	.841	.001

NS < .05 → Prédicteur significatif

VRAISEMBLANCE DES DONNÉES

Probabilité d'observer les données

$[(x_1, y_1), \dots, (x_p, y_i), \dots, (x_n, y_n)]$

$$\begin{aligned}
 &= \prod_{i=1}^n \text{Prob}(Y = y_i / X = x_i) \\
 &= \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \\
 &= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \sum_j \beta_j x_j}}{e^{\beta_0 + \sum_j \beta_j x_j} + 1} \right)^{y_i} \left(\frac{1}{e^{\beta_0 + \sum_j \beta_j x_j} + 1} \right)^{1-y_i} \\
 &= \ell(\beta_0, \beta_1, \dots, \beta_4)
 \end{aligned}$$

Le modèle de la régression logistique

Le modèle

$$\pi(x) = P(Y = F / X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_4 x_4}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_4 x_4}}$$

RÉSULTATS

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27.443	.543	.725

Collinearity Statistics	
Tolerance	VIF
X1	0.212
X2	0.252
X3	0.635
X4	0.904

Variables in the Equation

Step	B	S.E.	Wald	df	Sig.	Exp(B)
1	X1	6.002	1.414	1	.234	.001
	X2	3.703	.073	1	.786	40.581
	X3	-3.415	1.204	1	.005	.033
	X4	2.968	3.065	1	.333	19.461
	Constant	5.320	2.366	1	.025	204.283

a. Variable(s) entered on step 1: X1, X2, X3, X4.

RÉSULTATS

Correlations

	X1	X2	X3	X4
X1				
Pearson Correlation	1	.858**	.571**	-.053
Sig. (2-tailed)	.	.000	.000	.725
N	46	46	46	46
X2				
Pearson Correlation	.858**	1	.471**	.055
Sig. (2-tailed)	.000	.	.001	.717
N	46	46	46	46
X3				
Pearson Correlation	.571**	.471**	1	.154
Sig. (2-tailed)	.000	.001	.	.306
N	46	46	46	46
X4				
Pearson Correlation	-.053	.055	.154	1
Sig. (2-tailed)	.725	.717	.306	.
N	46	46	46	46

** . Correlation is significant at the 0.01 level (2-tailed).

Le modèle estimé

$$\text{Prob}(Y = F / X) =$$

$$\frac{e^{5.320 - 7.138 \times X_1 + 3.703 \times X_2 - 3.415 \times X_3 + 2.968 \times X_4}}{1 + e^{5.320 - 7.138 \times X_1 + 3.703 \times X_2 - 3.415 \times X_3 + 2.968 \times X_4}}$$

Prévision de faillite

Classification Table^a

Observed SITUATION	NF	Predicted		Percentage Correct
		FAILLITE	F	
NF	24		1	96.0
F	3		18	85.7
Overall Percentage				91.3

a. The cut value is .500

TEST DE HOSMER & LEMESHOW

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.201	7	.636

Contingency Table for Hosmer and Lemeshow Test

Step	FAILLITE = Non		FAILLITE = Oui		Total
	Observed	Expected	Observed	Expected	
1	5	4.999	0	.001	5
2	5	4.906	0	.094	5
3	4	4.613	1	.387	5
4	5	4.143	0	.857	5
5	4	3.473	1	1.527	5
6	1	1.762	4	3.238	5
7	0	.667	5	4.333	5
8	1	.340	4	4.660	5
9	0	.098	6	5.902	6

RÉGRESSION LOGISTIQUE PAS À PAS DESCENDANTE

Sans X₂

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27.516	.542	.724

Variables in the Equation

Step	B	S.E.	Wald	df	Sig.	Exp(B)	
1	X1	-5.772	3.005	3.690	1	.055	.003
	X3	-3.289	1.085	9.183	1	.002	.037
	X4	2.979	3.025	.970	1	.325	19.675
	Constant	5.038	2.060	5.983	1	.014	154.193

a. Variable(s) entered on step 1: X1, X3, X4.

RÉGRESSION LOGISTIQUE PAS À PAS DESCENDANTE

Sans X₄

Model Summary

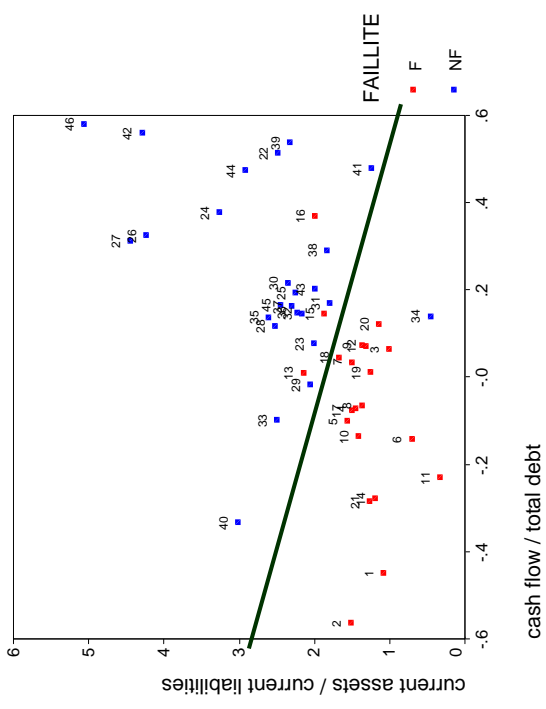
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	28.636	.531	.709

Variables in the Equation

Step	B	S.E.	Wald	df	Sig.	Exp(B)	
1	X1	-6.556	2.905	5.092	1	.024	.001
1	X3	-3.019	1.002	9.077	1	.003	.049
	Constant	5.940	1.986	8.950	1	.003	379.996

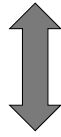
a. Variable(s) entered on step 1: X1, X3.

CARTE DES ENTREPRISES DANS LE PLAN (X₁, X₃)

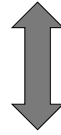


ÉQUATION DE LA DROITE FRONTIÈRE

$$\text{Prob}(Y = F / X) = \frac{e^{5.940 - 6.556 \times X_1 - 3.019 \times X_3}}{1 + e^{5.940 - 6.556 \times X_1 - 3.019 \times X_3}} = 0.5$$

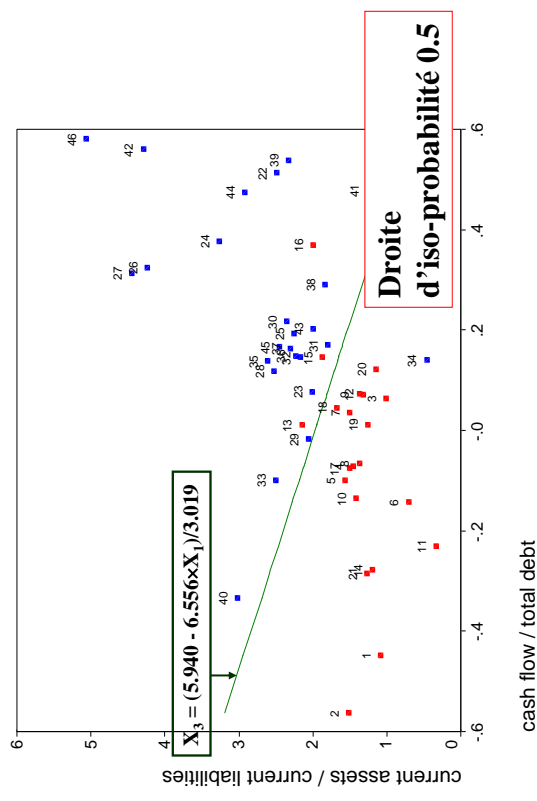


$$5.940 - 6.556 \times X_1 - 3.019 \times X_3 = 0$$



$$X_3 = (5.940 - 6.556 \times X_1) / 3.019$$

Carte des entreprises dans le plan (x₁, x₃) avec la droite frontière issue de la régression logistique



EXEMPLE II : LOW BIRTH WEIGHT BABY (HOSMER & LEMESHOW)

Y = 1 si le poids du bébé < 2 500 grammes, = 0 sinon
n₁ = 59, n₀ = 130

Facteurs de risque :

- Age
- LWT (Last Menstrual Period Weight)
- Race (White, Black, Other)
- FTV (Nb of First Trimester Physician Visits)
- Smoke (1 = YES, 0 = NO)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	214.575	.101	.142

Variables in the Equation

Step	AGE	LWT	WHITE	BLACK	FTV	SMOKE	Constant	B	S.E.	Wald	df	Sig.	Exp(B)
1								-.022	.035	.410	1	.522	.978
								-.012	.006	3.762	1	.052	.988
								-.941	.418	5.070	1	.024	.390
								.289	.527	.301	1	.583	1.336
								-.008	.164	.002	1	.962	.992
								1.053	.381	7.637	1	.006	2.866
								1.269	1.023	1.539	1	.215	3.558

a. Variable(s) entered on step 1: AGE, LWT, WHITE, BLACK, FTV, SMOKE.

Coefficients^a

Model	AGE	Collinearity Statistics	
		Tolerance	VIF
1	weight last menstrual period smoking during pregnancy n° physician visits first trimester	.884	1.132
	WHITE	.869	1.150
	BLACK	.865	1.156
		.939	1.065
		.686	1.457
		.743	1.346

a. Dependent Variable: low birth weight

Aucun problème de multicollinéarité

Validité du modèle Test de Hosmer et Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.825	8	.159

Contingency Table for Hosmer and Lemeshow Test

Step	low birth weight = weight > 2500 g		low birth weight = weight < 2500 g		Total
	Observed	Expected	Observed	Expected	
2	17	16.411	2	2.589	19
3	14	15.454	5	3.546	19
4	12	13.955	7	5.045	19
5	16	13.041	3	5.959	19
6	12	12.589	7	6.401	19
7	9	12.084	10	6.916	19
8	9	11.569	10	7.431	19
9	13	10.715	6	8.285	19
10	9	6.888	9	11.112	18

Résultats

ODDS-RATIO

Odds – Ratio(Smoke)

$$= \frac{\text{Pr ob}(Y = 1 / X, \text{Smoke} = \text{yes}) / \text{Pr ob}(Y = 0 / X, \text{Smoke} = \text{yes})}{\text{Pr ob}(Y = 1 / X, \text{Smoke} = \text{no}) / \text{Pr ob}(Y = 0 / X, \text{Smoke} = \text{no})} = \exp(\beta_{\text{Smoke}})$$

Pour un événement rare l'odds-ratio est peu différent du risque relatif défini par :

$$\text{Risque Relatif} = \frac{\text{Pr ob}(Y = 1 / X, \text{Smoke} = \text{yes})}{\text{Pr ob}(Y = 1 / X, \text{Smoke} = \text{no})}$$

INTERVALLE DE CONFIANCE DE L'ODDS-RATIO AU NIVEAU 95%

De

$$\text{Var}(\hat{\beta}_{\text{Smoke}}) = s_{\text{Smoke}}^2$$

on déduit l'intervalle de confiance de OR(Smoke) :

$$\left[e^{\hat{\beta}_{\text{Smoke}} - 1.96s_{\text{Smoke}}}, e^{\hat{\beta}_{\text{Smoke}} + 1.96s_{\text{Smoke}}} \right]$$

53

INFLUENCE D'UN GROUPE DE VARIABLES

Le modèle

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Test

$$\begin{aligned} H_0 : \beta_{r+1} = \dots = \beta_k = 0 \\ H_1 : \text{au moins un } \beta_j \neq 0 \end{aligned}$$

Statistiques utilisées

1. $\Lambda = [-2L(\text{Modèle simplifié})] - [-2L(\text{Modèle complet})]$

2. Wald = $(\hat{\beta}_{r+1}, \dots, \hat{\beta}_k) \left[\text{Var} \begin{pmatrix} \hat{\beta}_{r+1} \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \right]^{-1} \begin{pmatrix} \hat{\beta}_{r+1} \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$

- Proc Logistic
- Proc Genmod (type 3 et wald)
- SPSS

55

INTERVALLE DE CONFIANCE DE L'ODDS-RATIO AU NIVEAU 95%

Variables in the Equation

Step	B	Exp(B)	95.0% C.I. for EXP(B)	
			Lower	Upper
1	AGE	-.022	.978	1.047
	LWT	-.012	.988	1.000
	WHITE	-.941	.390	.885
	BLACK	.289	1.336	3.755
	FTV	-.008	.992	1.369
	SMOKE	1.053	2.866	6.046
	Constant	1.269	3.558	

a. Variable(s) entered on step 1: AGE, LWT, WHITE, BLACK, FTV, SMOKE.

54

RÈGLE DE DÉCISION

On rejette

$$H_0 : \beta_{r+1} = \dots = \beta_k = 0$$

au risque α de se tromper si

$$\Lambda \text{ ou Wald} \geq \chi_{1-\alpha}^2 [k - r]$$

ou si

$$\text{NS} = \text{Prob}(\chi^2 [k - r] \geq \text{Wald ou } \Lambda) \leq \alpha$$

55

TEST DU FACTEUR RACE (WALD)

Variables in the Equation

Step	B	S.E.	Wald	df	Sig.	Exp(B)
1						
AGE	-.022	.035	.410	1	.522	.978
LWT	-.012	.006	3.762	1	.052	.988
RACE			7.784	2	.020	
RACE(1)	-.941	.418	5.070	1	.024	.390
RACE(2)	.289	.527	.301	1	.583	1.336
SMOKE	1.053	.381	7.637	1	.006	2.866
FTV	-.008	.164	.002	1	.962	.992
Constant	1.269	1.023	1.539	1	.215	3.558

a. Variable(s) entered on step 1: AGE, LWT, RACE, SMOKE, FTV.

Modèle sans le facteur Race :

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	222.815	.061	.086

57

Test de l'hypothèse linéaire générale

Le modèle

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Test

$$\begin{aligned} H_0 : C(\beta_0, \beta_1, \dots, \beta_k)' &= 0 \\ H_1 : C(\beta_0, \beta_1, \dots, \beta_k)' &\neq 0 \end{aligned}$$

Statistiques utilisées

$$1. \Lambda = [-2L(H_0)] - [-2L(H_1)]$$

$$2. \text{Wald} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) C' [CVar(\begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}) C]^{-1} C \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Proc GENMOD

Proc Logistic
Proc Genmod

65

TEST DU FACTEUR RACE (LRT)

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	214.575 ^a	.000	0	.
AGE	214.990	.415	1	.520
LWT	218.746	4.171	1	.041
SMOKE	222.573	7.998	1	.005
FTV	214.577	.002	1	.963
RACE	222.815	8.239	2	.016

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

85

Règle de décision

On rejette

$$H_0 : C(\beta_0, \beta_1, \dots, \beta_k)' = 0$$

au risque α de se tromper si

$$\Lambda \text{ ou Wald} \geq \chi_{1-\alpha}^2 [\text{rang}(L)]$$

ou si

$$NS = \text{Prob}(\chi^2 [\text{rang}(L)] \geq \text{Wald ou } \Lambda) \leq \alpha$$

09

La régression logistique pas-à-pas descendante

- On part du modèle complet.
- A chaque étape, on enlève la variable ayant le Wald le moins significatif (plus fort niveau de signification) à condition que son niveau de signification soit supérieur à 10 % .

19

Test du Score pour la variable X_j

Modèle

$$\text{Prob}(Y = 1 / \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_t x_t + \beta_j x_j}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_t x_t + \beta_j x_j}}$$

Test $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

Statistique

$$\chi^2_{Score} = \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{t_0}}^{-1} \left[-\frac{\partial^2 L}{\partial \beta^2} \right]_{\beta = \hat{\beta}_{t_0}}^{-1} \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{t_0}}$$

suit une loi du khi-deux à 1 degré de liberté sous H_0 .

$\frac{\partial L}{\partial \beta}$ est calculé sur le modèle à $t+1$ variables.

29

Test du Score pour les variables hors modèle

Modèle

$$\text{Prob}(Y = 1 / \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_t x_t + \beta_{t+1} x_{t+1} + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_t x_t + \beta_{t+1} x_{t+1} + \dots + \beta_k x_k}}$$

Test $H_0 : \beta_{t+1} = \dots = \beta_k = 0$ vs $H_1 : \text{au moins un } \beta_j \neq 0$

Statistique

$$\chi^2_{Score} = \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{t_0}}^{-1} \left[-\frac{\partial^2 L}{\partial \beta \partial \beta'} \right]_{\beta = \hat{\beta}_{t_0}}^{-1} \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{t_0}}$$

suit une loi du khi-deux à $k-t$ degré de liberté sous H_0 .

$\frac{\partial L}{\partial \beta}$ est calculé sur le modèle à k variables.

39

RÉGRESSION LOGISTIQUE MULTIPLE (DONNÉES AGRÉGÉES)

Exemple : Job satisfaction (Models for discrete data, D. Zelterman, Oxford Science Publication, 1999)

9949 employees in the 'craft' job (travail manuel) within a company

Response : Satisfied/Dissatisfied

Factors : Sex (1=F, 0=M)

Race (White=1, Nonwhite=0)

Age (<35, 35-44, >44)

Region (Northeast, Mid-Atlantic, Southern, Midwest, Northwest, Southwest, Pacific)

Explain Job satisfaction with all the main effects and the interactions.

64

Job satisfaction (Y/N) by sex (M/F), race, age, and region of residence for employees of a large U.S. corporation

Region	White						Nonwhite					
	Under 35		35-44		Over 44		Under 35		35-44		Over 44	
	M	F	M	F	M	F	M	F	M	F	M	F
Northeast	288	60	224	35	337	70	38	19	32	22	21	15
Y	177	57	166	19	172	30	33	35	11	20	8	10
N												
Mid-Atlantic	90	19	96	12	124	17	18	13	7	0	9	1
Y	45	12	42	5	39	2	6	7	2	3	2	1
N												
Southern	226	88	189	44	156	70	45	47	18	13	11	9
Y	128	57	117	34	73	25	31	35	3	7	2	2
N												
Midwest	285	110	225	53	324	60	40	66	19	25	22	11
Y	179	93	141	24	140	47	25	56	11	19	2	12
N												
Northwest	270	176	215	80	269	110	36	25	9	11	16	4
Y	180	151	108	40	136	40	20	16	7	5	3	5
N												
Southwest	252	97	162	47	199	62	69	45	14	8	14	2
Y	126	61	72	27	93	24	27	36	7	4	5	0
N												
Pacific	119	62	66	20	67	25	45	22	15	10	8	6
Y	58	33	20	10	21	10	16	15	10	8	6	2
N												

UTILISATION DE LA PROC LOGISTIC

```
data job;
input sat nsat race age sex region;
label
    sat = 'satisfied with job'
    nsat = 'dissatisfied'
    race = '0=non-white, 1=white'
    age = '3 age groups'
    sex = '0=M, 1=F'
    region = '7 regions'
    total = 'denominator';
total = sat+nsat;
proprat = sat/total;
cards;
288 177 1 0 0 0
90 45 1 0 0 1
226 128 1 0 0 2
. .
. .
2 0 0 2 1 5
6 2 0 2 1 6
;
```

UTILISATION DE LA PROC LOGISTIC

```
proc logistic data=job;
class race age sex region/param=effect;
model sat/total = race age sex region race*age
race*sex race*region age*sex
age*region sex*region
/selection = forward
hierarchy = none ;
run;
```

RÉSULTAT DE LA PROC LOGISTIC (OPTION FORWARD ET HIERARCHY =NONE)

Type III Analysis of Effects

Effect	DF	Chi-Square	Wald	Pr > ChiSq
race	1	0.1007	0.7510	0.7510
age	2	50.7100	<.0001	<.0001
sex	1	14.0597	0.0002	0.0002
region	6	37.7010	<.0001	<.0001
race*sex	1	7.5641	0.0060	0.0060
age*sex	2	5.9577	0.0509	0.0509

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.6491	0.0346	350.2297	<.0001
race	0	-0.0099	0.0312	0.1007	0.7510
age	0	-0.1952	0.0316	38.2459	<.0001
age	1	-0.0227	0.0375	0.3675	0.5444
sex	0	0.1230	0.0328	14.0597	0.0002
region	0	-0.2192	0.0469	21.8470	<.0001
region	1	0.2228	0.0820	7.3832	0.0066
region	2	-0.0446	0.0527	0.7159	0.3975
region	3	-0.1291	0.0462	7.8133	0.0052
region	4	-0.0927	0.0472	3.8616	0.0494
region	5	0.0704	0.0531	1.7565	0.1851
race*sex	0 0	0.0856	0.0311	7.5641	0.0060
age*sex	0 0	0.0768	0.0315	5.9428	0.0148
age*sex	1 0	-0.0342	0.0375	0.8352	0.3608

```
proc logistic data=job;
class race age sex region/param=effect;
model sat/total = race age sex region
      race*sex age*sex ;
contrast 'Age >44' age -1 -1/estimate = parm;
contrast 'Pacific' region -1 -1 -1 -1 -1 -1/
      estimate=parm;
contrast 'Age>44,Homme' age*sex -1 -1/
      estimate=parm;
run;
```

Logit(Prob(Satisfait)) =

$$0.65 + \begin{matrix} \text{Non-Blanc} \\ \text{Blanc} \end{matrix} \begin{bmatrix} -.01 \\ +.01 \end{bmatrix} + \begin{matrix} < 35 \\ 35-44 \\ > 44 \end{matrix} \begin{bmatrix} -.20 \\ -.02 \\ +.22 \end{bmatrix} + \begin{matrix} \text{Homme} \\ \text{Femme} \end{matrix} \begin{bmatrix} +.12 \\ -.12 \end{bmatrix} + \begin{matrix} \text{Northeast} \\ \text{Mid-Atlantic} \\ \text{Southern} \\ \text{Midwest} \\ \text{Northwest} \\ \text{Southwest} \\ \text{Pacific} \end{matrix} \begin{bmatrix} -.22 \\ +.22 \\ -.04 \\ -.13 \\ -.09 \\ +.07 \\ +.19 \end{bmatrix}$$

$$\begin{matrix} \text{Non-Blanc} \\ \text{Blanc} \end{matrix} \begin{bmatrix} +.09 \\ -.09 \end{bmatrix} + \begin{matrix} < 35 \\ 35-44 \\ > 44 \end{matrix} \begin{bmatrix} +.08 \\ -.03 \\ -.05 \end{bmatrix} + \begin{matrix} \text{Homme} \\ \text{Femme} \end{matrix} \begin{bmatrix} -.08 \\ +.03 \\ +.05 \end{bmatrix}$$

Contrast Rows Estimation and Testing Results

Contrast	Estimate	Standard Error	Chi-Square	Wald	Pr > ChiSq
Age >44	0.2180	0.0375	0.1444		<.0001
Pacific	0.1924	0.0751	0.0453		0.0104
Age>44,Homme	-0.0425	0.0375	-0.1159		0.2565

Construction d'un modèle hiérarchique

```

proc logistic data=job;
class race age sex region/param=effect;
model sat/total= sex region race(sex)
age(sex) /scale=none ;
contrast 'Pacific' region -1 -1 -1 -1 -1 -1
/estimate=parm;
contrast 'Age>44,Homme' age(sex) -1 -1 0 0
/estimate = parm;
contrast 'Age>44,Femme' age(sex) 0 0 -1 -1 -1
/estimate=parm;
run;

```

73

RÉSULTATS

Type III Analysis of Effects

Effect	DF	Chi-Square	Wald	Pr > ChiSq
sex	1	14.0597	14.0597	0.0002
region	6	37.7010	37.7010	<.0001
race(sex)	2	7.5710	7.5710	0.0227
age(sex)	4	55.4078	55.4078	<.0001

74

RÉSULTATS

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.6481	0.0346	350.2297	<.0001
sex	1	0.1230	0.0328	14.0597	0.0002
region	0	-0.2192	0.0469	21.8470	<.0001
region	1	0.2228	0.0820	7.3832	0.0066
region	2	-0.0446	0.0527	0.7159	0.3975
region	3	-0.1291	0.0462	7.8133	0.0052
region	4	-0.0927	0.0472	3.8616	0.0494
region	5	0.0704	0.0531	1.7565	0.1851
race(sex)	0	0.0757	0.0422	3.2230	0.0726
race(sex)	0	-0.0956	0.0459	4.3244	0.0376
age(sex)	0	-0.1185	0.0342	11.9881	0.0005
age(sex)	1	-0.0570	0.0370	2.3683	0.1238
age(sex)	0	-0.2720	0.0530	26.3735	<.0001
age(sex)	1	0.0115	0.0652	0.0313	0.8596

Contrast	Estimate	Standard Error	Chi-Square	Wald	Pr > ChiSq
Pacific	0.1924	0.0751	6.5729	6.5729	0.0104
Age>44,Homme	0.1754	0.0367	22.8477	22.8477	<.0001
Age>44,Femme	0.2605	0.0654	15.8719	15.8719	<.0001

75

Utilisation de la Proc Logistic avec l'option Param=effect

$$\text{Logit}(\text{Prob}(\text{Satisfait})) =$$

$$0.65 + \begin{matrix} \text{Homme} & \begin{bmatrix} +.12 \\ - .12 \end{bmatrix} \\ \text{Femme} & \end{matrix} + \begin{matrix} \text{Northeast} & - .22 \\ \text{Mid - Atlantic} & + .22 \\ \text{Southern} & - .04 \\ \text{Midwest} & - .13 \\ \text{Northwest} & - .09 \\ \text{Southwest} & + .07 \\ \text{Pacific} & + .19 \end{matrix} \begin{matrix} \text{ns} \\ \\ \\ \\ \text{ns} \\ \end{matrix}$$

$$+ \begin{matrix} \text{Homme} & \begin{bmatrix} +.08 & -.08 \\ -.10 & +.10 \end{bmatrix} \\ \text{Femme} & \end{matrix} \begin{matrix} \text{Non-blanc} & \text{Blanc} \end{matrix}$$

$$+ \begin{matrix} \text{Homme} & \begin{bmatrix} -.12 & -.06 \\ -.27 & .01 \end{bmatrix} \\ \text{Femme} & \end{matrix} \begin{matrix} >35 & 35-44 & >44 \end{matrix} \begin{matrix} \text{ns} \\ \text{ns} \\ \end{matrix}$$

Différence entre races par sexe : Race(Sexe)

Différence entre les ages par sexe : Age(Sexe)

76

Analyse des résidus données agrégées en s groupes

- n_i = effectif du groupe i , $i = 1$ à $s = 84$
- y_i = nombre de succès observé dans le groupe i
- $\hat{\pi}_i$ = probabilité de succès dans le groupe i
- $\hat{y}_i = n_i \hat{\pi}_i$ = nombre de succès attendu dans le groupe i

- Résidu de Pearson :

$$r_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

- Résidu déviance :

$$d_i = \text{signe}(y_i - \hat{y}_i) \sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)}$$

77

Analyse des résidus et validation du modèle

```
proc logistic data=job;
class race age sex region/param=effect;
model sat/total=race age sex region
      race*sex age*sex / scale = none ;
output out = residu
       predicted =predicted
       reschi =reschi   resdev=resdev;
run;

Proc print data=residu;
var sat total propsat predicted reschi resdev;
run;
```

8L

Analyse des résidus : Résultats

Obs	sat	total	propsat	predicted	reschi	resdev
1	288	465	0.61935	0.58848	1.35305	1.35864
2	90	135	0.66667	0.68991	-0.58388	-0.58005
3	226	354	0.63842	0.63003	0.32704	0.32756
4	285	464	0.61422	0.61011	0.18152	0.18164
5	270	450	0.60000	0.61875	-0.81897	-0.81651
6	252	378	0.66667	0.65641	0.41995	0.42097
7	119	177	0.67232	0.68338	-0.31638	-0.31541
8	60	117	0.51282	0.53231	-0.42246	-0.42216
9	19	31	0.61290	0.63909	-0.30364	-0.30214

79

Validation du modèle

- Le khi-deux de Pearson :

$$Q_P = \sum_{i=1}^s r_i^2$$

- La déviance :

$$Q_L = \sum_{i=1}^s d_i^2$$

- Si le modèle étudié est exact Q_P et Q_L suivent approximativement une loi du khi-deux à [nb de groupes - nb de paramètres du modèle] degrés de liberté.

08

Remarques

- Les tests de validation sont valables s'il y a au moins 10 sujets par groupe.
- La déviance Q_L est égale à

$$[-2L(\text{modèle étudié}) - [-2L(\text{modèle saturé})]]$$

où le modèle saturé est un modèle reconstituant parfaitement les données.

81

Résultats

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	70	81.9676	1.1710	0.1552
Pearson	70	79.0760	1.1297	0.2142

Number of events/trials observations: 84

82

SUR-DISPERSION

Khi-deux de Pearson Q_p et déviance Q_L sont trop forts si :

- Modèle mal spécifié
- Outliers

Hétérogénéité de chaque groupe

La variable de réponse $Y_i = \text{Nb de succès sur le groupe } i$ ne suit plus une loi binomiale :

- $E(Y_i) = n_i \pi_i$
- $V(Y_i) = \phi n_i \pi_i (1 - \pi_i)$

83

CALCUL DE ϕ

Dans la Proc LOGISTIC :

- Option SCALE = Pearson :
- Option SCALE = Deviance :

$$\phi = \frac{Q_p}{ddl}$$

$$\phi = \frac{Q_L}{ddl}$$

Dans la Proc GENMOD :

- Option PSCALE ou DSCALE
- Scale = (vrai également dans Proc Logistic)

$$\sqrt{\phi}$$

84

SOLUTION LOGISTIC/GENMOD POUR PRENDRE EN COMPTE LA SUR-DISPERSION

Utilisation de la réponse binomiale pour l'estimation des paramètres.

Pour les tests sur les coefficients :

- Les statistiques de Wald et LRT sont divisées par ϕ .
- Les déviations sont divisées par ϕ .
- Dans GENMOD, utilisation de la statistique

$$F = \frac{Dev(\text{Modèle sous } H_0) - Dev(\text{Modèle sous } H_1)}{(ddl_{H_0} - ddl_{H_1}) \times \phi}$$

S'il y a sur-dispersion (Déviance et Khi-deux de Pearson significatifs) les résultats non corrigés sont trop significatifs.

B. LA RÉGRESSION LOGISTIQUE ORDINALE

Exemple : Qualité des vins de Bordeaux

Variables observées sur 34 années (1924 - 1957)

TEMPERATURE : Somme des températures moyennes journalières

SOLEIL : Durée d'insolation

CHALEUR : Nombre de jours de grande chaleur

PLUIE : Hauteur des pluies

QUALITE DU VIN : **Bon, Moyen, Médiocre**

LES DONNÉES

	Température	Soleil	Chaleur	Pluie	Qualité
1	3064	1201	10	361	2
2	3000	1053	11	338	3
3	3155	1133	19	393	2
4	3085	970	4	467	3
5	3245	1258	36	294	1
6	3267	1386	35	225	1
7	3080	966	13	417	3
8	2974	1189	12	488	3
9	3038	1103	14	677	3
10	3318	1310	29	427	2
11	3317	1362	25	326	1
12	3182	1171	28	326	3
13	2998	1102	9	349	3
14	3221	1424	21	382	1
15	3019	1230	16	275	2
16	3022	1285	9	303	2
17	3084	1329	11	339	2
18	3009	1210	15	536	3
19	3227	1331	21	414	2
20	3308	1366	24	282	1
21	3212	1289	17	302	2
22	3361	1444	25	253	1
23	3061	1175	12	261	2
24	3478	1317	42	259	1
25	3126	1248	11	315	2
26	3458	1508	43	286	1
27	3252	1361	26	346	2
28	3052	1186	14	443	3
29	3270	1399	24	306	1
30	3198	1259	20	367	1
31	2904	1164	6	311	3
32	3247	1277	19	375	1
33	3083	1195	5	441	3
34	3043	1208	14	371	3

Correlations					
Température	Pearson Correlation	Température	Soleil	Chaleur	Pluie
		1			
	Sig. (2-tailed)		.712**	.865**	-.410*
	N	34	34	34	34
Soleil	Pearson Correlation	.712**	1	.646**	-.473**
	Sig. (2-tailed)	.000		.000	.005
	N	34	34	34	34
Chaleur	Pearson Correlation	.865**	.646**	1	-.401*
	Sig. (2-tailed)	.000	.000		.019
	N	34	34	34	34
Pluie	Pearson Correlation	-.410*	-.473**	-.401*	1
	Sig. (2-tailed)	.016	.005	.019	
	N	34	34	34	34

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

CORRÉLATIONS

Coefficients^a

Model	Temperature Soleil Chaleur Pluie	Collinearity Statistics	
		Tolerance	VIF
1	Temperature Soleil Chaleur Pluie	.211 .451 .248 .760	4.733 2.216 4.031 1.316

a. Dependent Variable: Qualité

VIF

LA RÉGRESSION LOGISTIQUE ORDINALE

La variable Y prend 1, ..., m, m+1 valeurs ordonnées.

I. Le modèle à pentes égales

Dans la Proc Logistic :

$$\text{Prob}(Y \leq i / X) = \frac{e^{\alpha_i + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\alpha_i + \beta_1 x_1 + \dots + \beta_k x_k}}$$

pour $i = 1, \dots, m$ et avec $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_m$

Dans SPSS :

$$\text{Prob}(Y \leq i / X) = \frac{e^{\alpha_i - \theta_1 x_1 - \dots - \theta_k x_k}}{1 + e^{\alpha_i - \theta_1 x_1 - \dots - \theta_k x_k}}$$

Les coefficients de régression des x_j de SPSS sont l'opposé de ceux de SAS : $\theta_j = -\beta_j$.

PROPRIÉTÉS DU MODÈLE

Modèle à pentes égales (proportional odds ratio)

$$\frac{\text{Prob}(Y \leq i/x) / \text{Prob}(Y > i/x)}{\text{Prob}(Y \leq i/x') / \text{Prob}(Y > i/x')} = \frac{e^{\alpha_i + x\beta}}{e^{\alpha_i + x'\beta}} = e^{(x-x')\beta}$$

est indépendant de i.

Lorsque $\beta_j > 0$, la probabilité des petites valeurs de Y augmente avec X_j .

TEST DU MODÈLE À PENTES ÉGALES DANS SAS

Le modèle général

$$\text{Prob}(Y \leq i / X) = \frac{e^{\alpha_i + \beta_{11}x_1 + \dots + \beta_{k1}x_k}}{1 + e^{\alpha_i + \beta_{11}x_1 + \dots + \beta_{k1}x_k}}$$

pour $i = 1, \dots, m$

Test H_0 :

$$\left. \begin{aligned} \beta_{11} = \beta_{12} = \dots = \beta_{1m} \\ \beta_{21} = \beta_{22} = \dots = \beta_{2m} \\ \vdots \\ \beta_{k1} = \beta_{k2} = \dots = \beta_{km} \end{aligned} \right\} k(m-1) \text{ contraintes}$$

STATISTIQUE UTILISÉE

$L(\beta) =$ Log-vraisemblance du modèle général
 $\beta_{H_0} =$ estimation de β pour le modèle à pentes égales

La statistique

$$\chi^2_{Score} = \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{H_0}} \left[- \frac{\partial^2 L}{\partial \beta \partial \beta'} \right]_{\beta = \hat{\beta}_{H_0}}^{-1} \left[\frac{\partial L}{\partial \beta} \right]_{\beta = \hat{\beta}_{H_0}}$$

suit une loi du khi-deux à $k(m-1)$ degrés de liberté sous l'hypothèse H_0 .

RÈGLE DE DÉCISION

On rejette l'hypothèse H_0 d'un modèle à pentes égales au risque α de se tromper si

$$\chi^2_{Score} \geq \chi^2_{1-\alpha} [k(m-1)]$$

ou si

$$NS = \text{Prob}(\chi^2 [m(k-1)] \geq \chi^2_{Score}) \leq \alpha$$

Conseil d'Agresti :

Test plutôt utilisé pour valider H_0 que pour rejeter H_0 .

RÉSULTATS SPSS

Test of Parallel Lines^a

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis General	26.158	3.803	4	.433

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	74.647			
Final	26.158	48.489	4	.000

Link function: Logit.

Pseudo R-Square

Cox and Snell	.760
Nagelkerke	.855
McFadden	.650

Link function: Logit.

RÉSULTATS SPSS

Modèle complet

Parameter Estimates

	Estimate	Std. Error	Wald	df	Sig.
Threshold [QUALITE = 1]	-.8550748	34.921740	5.99549	1	.014
[QUALITE = 2]	-.8054960	33.96555	5.62405	1	.018
TEMPERAT	-.02427	.01277	3.61247	1	.057
SOLEIL	-.01379	.00850	2.63346	1	.105
CHALEUR	.08876	.11929	.55364	1	.457
PLUIE	.02589	.01235	4.39307	1	.036

Link function: Logit.

Modèle sans Chaleur

Parameter Estimates

	Estimate	Std. Error	Wald	df	Sig.
Threshold [QUALITE = 1]	-.6744675	22.89023	8.66204	1	.003
[QUALITE = 2]	-.6263810	21.78872	8.26445	1	.004
TEMPERAT	-.01717	.00759	5.11905	1	.024
SOLEIL	-.01499	.00832	3.24843	1	.071
PLUIE	.02224	.01046	4.52311	1	.033

Link function: Logit.

Case Summaries^a

	Qualité	Estimated Cell Probability for Response Category 1	Estimated Cell Probability for Response Category 2	Estimated Cell Probability for Response Category 3	Predicted Response Category
1	Moyen	.01	.48	.51	Médiocre
2	Médiocre	.00	.05	.95	Médiocre
3	Moyen	.01	.44	.56	Médiocre
4	Médiocre	.00	.00	1.00	Médiocre
5	Bon	.64	.35	.00	Bon
6	Bon	.99	.01	.00	Bon
7	Médiocre	.00	.01	.99	Médiocre
8	Médiocre	.00	.01	1.00	Médiocre
9	Moyen	.42	.57	.00	Moyen
10	Bon	.94	.06	.00	Bon
11	Médiocre	.08	.83	.09	Médiocre
12	Médiocre	.07	.87	.06	Médiocre
13	Bon	.67	.33	.00	Bon
14	Moyen	.04	.78	.18	Moyen
15	Moyen	.05	.81	.15	Moyen
16	Moyen	.03	.82	.15	Moyen
17	Médiocre	.00	.00	.99	Médiocre
18	Moyen	.21	.76	.03	Moyen
19	Bon	.97	.03	.00	Bon
20	Moyen	.58	.42	.00	Moyen
21	Bon	1.00	.00	.00	Bon
22	Moyen	.04	.81	.15	Moyen
23	Bon	1.00	.00	.00	Bon
24	Bon	1.00	.00	.00	Bon
25	Moyen	.11	.83	.06	Moyen
26	Bon	1.00	.00	.00	Bon
27	Moyen	.75	.25	.00	Moyen
28	Médiocre	.00	.05	.95	Médiocre
29	Bon	.14	.81	.05	Bon
30	Bon	.00	.09	.90	Moyen
31	Médiocre	.00	.09	.90	Médiocre
32	Bon	.29	.69	.02	Moyen
33	Médiocre	.00	.17	.83	Médiocre
34	Médiocre	.00	.36	.63	Médiocre

a. Limited to first 100 cases.

PRÉVISION DE LA QUALITÉ DU VIN AVEC LE 2E MODÈLE

QUALITÉ DE LA PRÉVISION

Qualité * Predicted Response Category Crosstabulation

Count	Qualité	Predicted Response Category			Total
		Bon	Moyen	Médiocre	
	Bon	9	2		11
	Moyen	2	7	2	11
	Médiocre		1	11	12
	Total	11	10	13	34

II. LE MODÈLE PARTIELLEMENT À PENTES ÉGALES

Les données de chaque observation sont répétées m fois.

La variable « Type » indique le numéro de la répétition i.

La variable « Réponse » indique si [Y ≤ i] est vrai :

Année	Qualité	Type	Réponse
1926	2	1	0
1926	2	2	1
1927	3	1	0
1927	3	2	0
1928	1	1	1
1928	1	2	1

(Y=1) faux ←
(Y≤ 2) vrai ←

Pour Type = 1 : Réponse = 1 ⇔ Qualité = 1

Pour Type = 2 : Réponse = 1 ⇔ Qualité ≤ 2

LE MODÈLE COMPLET

Prob(Réponse = 1 / Type, x)

$$= \frac{e^{\alpha_1 T_1 + \alpha_2 T_2 + \beta_1 T + \dots + \beta_4 P + \beta_5 T_1 \times T + \dots + \beta_8 T_1 \times P}}{1 + e^{\alpha_1 T_1 + \alpha_2 T_2 + \beta_1 T + \dots + \beta_4 P + \beta_5 T_1 \times T + \dots + \beta_8 T_1 \times P}}$$

- Pour Type = 1 : Réponse = 1 ⇔ Qualité = 1
- Pour Type = 2 : Réponse = 1 ⇔ Qualité ≤ 2
- D'où : Prob(Réponse = 1/Type = 1, x) = Prob(Qualité = 1/x)
- Prob(Réponse = 1/Type = 2, x) = Prob(Qualité ≤ 2/x)

- T₁, T₂ = variables indicatrices de la variable Type

LE CODE SAS

```

Proc genmod data=bordeaux2 descending;
class type annee;
model reponse = type tempera soleil chaleur pluie
type*tempera type*soleil
type*chaleur type*pluie
/dist=bin link=logit type3 noint;
repeated subject=annee / type=unstr;
run;
    
```

RÉSULTATS ÉTAPE 1

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	58	22.5317	0.3885
Scaled Deviance	58	22.5317	0.3885
Pearson Chi-Square	58	20.4541	0.3527
Scaled Pearson X2	58	20.4541	0.3527
Log Likelihood		-11.2659	

Algorithm converged.

RÉSULTATS ÉTAPE 1

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits	Z	Pr > Z
Intercept	0.0000	0.0000	0.0000 0.0000	.	.
type 1	-68.1364	29.7166	-126.380 -9.8929	-2.29	0.0219
type 2	-251.965	82.1239	-412.925 -91.0055	-3.07	0.0022
tempera	0.0948	0.0330	0.0300 0.1596	2.87	0.0041
soleil	0.0079	0.0107	-0.0130 0.0288	0.74	0.4598
chaleur	-0.8727	0.3574	-1.5732 -0.1722	-2.44	0.0146
pluie	-0.1036	0.0437	-0.1893 -0.0179	-2.37	0.0178
tempera*type 1	-0.0755	0.0358	-0.1458 -0.0053	-2.11	0.0351
tempera*type 2	0.0000	0.0000	0.0000 0.0000	.	.
soleil*type 1	0.0013	0.0144	-0.0270 0.0295	0.09	0.9290
soleil*type 2	0.0000	0.0000	0.0000 0.0000	.	.
chaleur*type 1	0.8799	0.3795	0.1360 1.6238	2.32	0.0204
chaleur*type 2	0.0000	0.0000	0.0000 0.0000	.	.
pluie*type 1	0.0852	0.0460	-0.0050 0.1753	1.85	0.0641
pluie*type 2	0.0000	0.0000	0.0000 0.0000	.	.

RÉSULTATS

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
type	2	7.08	0.0290
tempera	1	4.94	0.0263
soleil	0	.	.
chaleur	2	0.00	0.9995
pluie	2	0.02	0.9881
tempera*type	2	0.04	0.9799
soleil*type	2	0.27	0.8734
chaleur*type	2	0.00	0.9999
pluie*type	2	0.00	1.0000

LE MODÈLE PARTIELLEMENT À PENTES ÉGALES

On élimine progressivement les interactions non significatives.

On retrouve le modèle à pentes égales si toutes les interactions sont éliminées.

Cette approche permet un test LRT de comparaison entre le modèle complet et le modèle à pentes égales.

RÉSULTAT DES ITÉRATIONS MODÈLE À PENTES ÉGALES

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	62	26.2408	0.4232
Scaled Deviance	62	26.2408	0.4232
Pearson Chi-Square	62	26.5218	0.4278
Scaled Pearson X2	62	26.5218	0.4278
Log Likelihood		-13.1204	

Algorithm converged.

RÉSULTAT DES ITÉRATIONS MODÈLE À PENTES ÉGALES

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000		
type	1	-86.4800	35.0585	-155.193 -17.7666	6.08	0.0196
tempera	2	-81.5119	34.0447	-148.238 -14.7855	5.73	0.0167
soleil	1	0.0245	0.0127	-0.0004 0.0495	3.70	0.0543
chaleur	1	0.0140	0.0085	-0.0026 0.0306	2.73	0.0986
pluie	1	-0.0922	0.1180	-0.3235 0.1391	0.61	0.4348
	1	-0.0259	0.0123	-0.0500 -0.0019	4.46	0.0347

C. RÉGRESSION LOGISTIQUE MULTINOMIALE

La variable nominale Y prend r valeurs.

Modèle : (La modalité r sert de référence.)

$$\text{Prob}(Y = i / x) = \frac{e^{\alpha_i + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}{1 + \sum_{i=1}^{r-1} e^{\alpha_i + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}, \quad i = 1, \dots, r-1$$

$$\text{Prob}(Y = r / x) = \frac{1}{1 + \sum_{i=1}^{r-1} e^{\alpha_i + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}$$

APPLICATION AUX VINS DE BORDEAUX LE CODE SAS

```
proc catmod data=bordeaux;
direct tempera soleil chaleur pluie;
response logit;
model qualite = tempera soleil chaleur pluie;
run;
```

Test de Wald sur l'influence d'une variable X_j

Le modèle

$$\pi_i(x) = P(Y = i / X = x) = \frac{e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}{1 + e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}, \quad i = 1, \dots, r-1$$

Test

$$\begin{aligned} H_0 : \beta_{ij} &= \dots = \beta_{r-1,j} = 0 \\ H_1 : &\text{au moins un } \beta_{ij} \neq 0 \end{aligned}$$

Statistique utilisée

$$\text{Wald} = (\hat{\beta}_{1j}, \dots, \hat{\beta}_{r-1,j}) \left[\text{Var} \begin{bmatrix} \hat{\beta}_{1j} \\ \vdots \\ \hat{\beta}_{r-1,j} \end{bmatrix} \right]^{-1} \begin{bmatrix} \hat{\beta}_{1j} \\ \vdots \\ \hat{\beta}_{r-1,j} \end{bmatrix}$$

601

111

Règle de décision

On rejette

$$H_0 : \beta_{1j} = \dots = \beta_{r-1,j} = 0$$

au risque α de se tromper si

$$\text{Wald} \geq \chi^2_{1-\alpha}[r-1]$$

ou si

$$\text{NS} = \text{Prob}(\chi^2[r-1] \geq \text{Wald}) \leq \alpha$$

112

Test de Wald sur l'influence d'une variable X_j

Le modèle

$$\pi_i(x) = P(Y = i / X = x) = \frac{e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}{1 + e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}, \quad i = 1, \dots, r-1$$

Test

$$\begin{aligned} H_0 : \beta_{ij} &= \dots = \beta_{r-1,j} = 0 \\ H_1 : &\text{au moins un } \beta_{ij} \neq 0 \end{aligned}$$

Statistique utilisée

$$\text{Wald} = (\hat{\beta}_{1j}, \dots, \hat{\beta}_{r-1,j}) \left[\text{Var} \begin{bmatrix} \hat{\beta}_{1j} \\ \vdots \\ \hat{\beta}_{r-1,j} \end{bmatrix} \right]^{-1} \begin{bmatrix} \hat{\beta}_{1j} \\ \vdots \\ \hat{\beta}_{r-1,j} \end{bmatrix}$$

601

111

Influence des p variables X_{p+1}, \dots, X_k

Le modèle

$$\pi_i(x) = P(Y = i / X = x) = \frac{e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}{1 + e^{\beta_{i0} + \beta_{i1}x_1 + \dots + \beta_{ik}x_k}}, \quad i = 1, \dots, r-1$$

Test

$$\begin{aligned} H_0 : \beta_{i,p+1} = \dots = \beta_{ik} &= 0, \quad i = 1, \dots, r-1 \\ H_1 : &\text{au moins un } \beta_{ij} \neq 0 \end{aligned}$$

Statistiques utilisées

1. $\Lambda = [-2L(\text{Modèle simplifié})] - [-2L(\text{Modèle complet})]$

2.
$$\text{Wald} = (\hat{\beta}_{1,p+1}, \dots, \hat{\beta}_{r-1,k}) \left[\text{Var} \begin{bmatrix} \hat{\beta}_{1,p+1} \\ \vdots \\ \hat{\beta}_{r-1,k} \end{bmatrix} \right]^{-1} \begin{bmatrix} \hat{\beta}_{1,p+1} \\ \vdots \\ \hat{\beta}_{r-1,k} \end{bmatrix}$$

112

112

APPLICATION AUX VINS DE BORDEAUX

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	74.647			
Final	22.227	52.420	8	.000

Pseudo R-Square

Cox and Snell	.786
Nagelkerke	.884
McFadden	.702

113

APPLICATION AUX VINS DE BORDEAUX

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	34.575	12.348	2	.002
TEMPERAT	29.546	7.319	2	.026
SOLEIL	22.870	.642	2	.725
CHALEUR	25.894	3.667	2	.160
PLUIE	31.242	9.015	2	.011

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Les tests LRT sont plus justes que les tests de Wald : meilleure approximation du niveau de signification.

114

APPLICATION AUX VINS DE BORDEAUX

Parameter Estimates

Qualité	Intercept	B	Std. Error	Wald	df	Sig.	Exp(B)
Bon							
Intercept	-313.557		230.325	1.853	1	.173	
TEMPERAT	.113		.096	1.375	1	.241	1.120
SOLEIL	.015		.024	.370	1	.543	1.015
CHALEUR	-.874		.934	.876	1	.349	.417
PLUIE	-.122		.104	1.387	1	.239	.885
Moyen							
Intercept	-249.604		225.594	1.224	1	.269	
TEMPERAT	.095		.095	.999	1	.318	1.099
SOLEIL	.007		.022	.094	1	.759	1.007
CHALEUR	-.890		.923	.930	1	.335	.411
PLUIE	-.105		.103	1.040	1	.308	.901

115

APPLICATION AUX VINS DE BORDEAUX

Pseudo R-Square

Cox and Snell	.782
Nagelkerke	.880
McFadden	.694

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	42.197	19.327	2	.000
TEMPERAT	43.392	20.522	2	.000
CHALEUR	30.419	7.550	2	.023
PLUIE	41.634	18.764	2	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

116

APPLICATION AUX VINS DE BORDEAUX

Parameter Estimates

Qualité	B	Std. Error	Wald	df	Sig.	Exp(B)
Bon						
Intercept	-381.353	190.219	4.019	1	.045	
TEMPERAT	.145	.074	3.893	1	.048	1.156
CHALEUR	-1.161	.738	2.478	1	.115	.313
PLUIE	-.151	.080	3.577	1	.059	.860
Moyen						
Intercept	-308.897	186.257	2.750	1	.097	
TEMPERAT	.121	.072	2.785	1	.095	1.129
CHALEUR	-1.145	.729	2.471	1	.116	.318
PLUIE	-.133	.078	2.906	1	.088	.875

APPLICATION AUX VINS DE BORDEAUX

Classification

	Bon	Predicted		Percent Correct
		Moyen	Médiocre	
Observed Bon	8	3	0	72.7%
Moyen	3	7	1	63.6%
Médiocre	0	1	11	91.7%
Overall Percentage	32.4%	32.4%	35.3%	76.5%

Case Summaries^a

	Qualité	Estimated Cell Probability for Response Category 1	Estimated Cell Probability for Response Category 2	Estimated Cell Probability for Response Category 3	Predicted Response Category
1	Moyen	.01	.88	.10	Moyen
2	Médiocre	.00	.03	.97	Médiocre
3	Moyen	.01	.19	.79	Médiocre
4	Médiocre	.00	.07	.93	Médiocre
5	Bon	.73	.26	.01	Bon
6	Bon	.94	.06	.00	Bon
7	Moyen	.00	.00	1.00	Médiocre
8	Médiocre	.00	.00	1.00	Médiocre
9	Moyen	.63	.34	.03	Bon
10	Bon	.92	.08	.00	Bon
11	Médiocre	.20	.50	.30	Moyen
12	Médiocre	.00	.04	.96	Médiocre
13	Bon	.30	.69	.01	Moyen
14	Moyen	.02	.85	.13	Moyen
15	Moyen	.02	.77	.21	Moyen
16	Moyen	.05	.85	.10	Moyen
17	Moyen	.00	.00	1.00	Médiocre
18	Moyen	.21	.72	.08	Moyen
19	Bon	.95	.05	.00	Bon
20	Bon	.60	.40	.00	Bon
21	Bon	.99	.01	.00	Bon
22	Bon	.08	.92	.00	Moyen
23	Bon	1.00	.00	.00	Bon
24	Bon	.14	.86	.00	Moyen
25	Bon	1.00	.00	.00	Bon
26	Bon	.62	.38	.00	Bon
27	Moyen	.00	.00	1.00	Médiocre
28	Bon	.84	.16	.00	Bon
29	Bon	.25	.75	.00	Moyen
30	Bon	.00	.00	1.00	Médiocre
31	Bon	.49	.51	.00	Moyen
32	Bon	.00	.38	.62	Médiocre
33	Médiocre	.00	.00	1.00	Médiocre
34	Médiocre	.00	.00	1.00	Médiocre

a. Limited to first 100 cases.

EXEMPLE ALLIGATORS (AGRESTI)

Table 1: Primary Food Choice of Alligators, by Lake, Gender, and Size

Lake	Gender	Size	Primary Food Choice				
			Fish	Invertebrate	Reptile	Bird	Other
Hancock	Male	≤2.3	7	1	0	0	5
	Female	>2.3	4	0	0	1	2
Oklawaha	Male	≤2.3	16	3	2	2	3
	Female	>2.3	3	0	1	2	3
Trafford	Male	≤2.3	2	2	0	0	1
	Female	>2.3	13	7	6	0	0
George	Male	≤2.3	3	9	1	0	2
	Female	>2.3	0	1	0	1	0
George	Male	≤2.3	3	7	1	0	1
	Female	>2.3	8	6	6	3	5
George	Male	≤2.3	2	4	1	1	4
	Female	>2.3	0	1	0	0	0
George	Male	≤2.3	13	10	0	2	2
	Female	>2.3	9	0	0	1	2
George	Male	≤2.3	3	9	1	0	1
	Female	>2.3	8	1	0	0	1

EXEMPLE ALLIGATORS

The sample consisted of 219 alligators captured in four Florida lakes, during September 1985.

The response variable is the primary food type, in volume, found in an alligator's stomach. This variable had five categories: Fish, Invertebrate, Reptile, Bird, Other.

The invertebrates found in the stomachs were primarily apple snails, aquatic insects, and crayfish.

The reptiles were primarily turtles (though one stomach contained tags of 23 baby alligators that had been released in the lake during the previous year!).

The Other category consisted of amphibian, mammal, plant material, stones or other debris, or no food of dominant type.

EXEMPLE ALLIGATORS

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	146.644 ^a	.000	0	.
LAKE	196.962	50.318	12	.000
GENDER	148.859	2.215	4	.696
SIZE	164.244	17.600	4	.001

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

EXEMPLE ALLIGATORS

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	95.028 ^a	.000	0	.
LAKE	144.161	49.133	12	.000
SIZE	116.115	21.087	4	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Modèle estimé

Parameter Estimates

CHOICE	B	Std. Error	Wald	df	Sig.
Intercept	-.626	.642	.952	1	.329
LAKE=G]	1.847	1.317	1.967	1	.161
LAKE=H]	1.300	.993	1.712	1	.191
LAKE=O]	-1.265	1.233	1.052	1	.305
LAKE=T]	0 ^a			0	
SIZE=<=2.3]	-.279	.806	.120	1	.729
SIZE=>2.3]	0 ^a			0	
Intercept	.379	.479	.626	1	.429
LAKE=G]	2.935	1.116	6.913	1	.009
LAKE=H]	1.692	.780	4.703	1	.030
LAKE=O]	.476	.634	.564	1	.452
LAKE=T]	0 ^a			0	
SIZE=<=2.3]	.351	.680	.367	1	.545
SIZE=>2.3]	0 ^a			0	
Intercept	-.048	.505	.009	1	.925
LAKE=G]	1.813	1.127	2.590	1	.108
LAKE=H]	-1.088	.908	1.434	1	.231
LAKE=O]	.292	.641	.207	1	.649
LAKE=T]	0 ^a			0	
SIZE=<=2.3]	1.809	.603	9.008	1	.003
SIZE=>2.3]	0 ^a			0	
Intercept	-.009	.522	.000	1	.987
LAKE=G]	1.419	1.189	1.424	1	.233
LAKE=H]	1.002	.830	1.459	1	.227
LAKE=O]	-1.034	.840	1.515	1	.218
LAKE=T]	0 ^a			0	
SIZE=<=2.3]	.683	.651	1.099	1	.295
SIZE=>2.3]	0 ^a			0	

a. This parameter is set to zero because it is redundant.

Prévision

Case Summaries

	LAKE	SIZE	Estimated Cell Probability for Response Category: B	Estimated Cell Probability for Response Category: F	Estimated Cell Probability for Response Category: I	Estimated Cell Probability for Response Category: O	Estimated Cell Probability for Response Category: R
1	H	<=2.3	.07	.54	.09	.25	.05
2	H	>2.3	.14	.57	.02	.19	.07
3	O	<=2.3	.01	.26	.60	.05	.08
4	O	>2.3	.03	.46	.25	.07	.19
5	T	<=2.3	.04	.18	.52	.17	.09
6	T	>2.3	.11	.30	.19	.20	.20
7	G	<=2.3	.03	.45	.41	.09	.01
8	G	>2.3	.08	.66	.14	.10	.02

H = Hancock, O = Oklawaha, T = Trafford, G = George

Exemple Alligators (2)

SEX	LENGTH	CHOICE	SEX	LENGTH	CHOICE	SEX	LENGTH	CHOICE
M	1.30	I	M	2.03	F	F	1.78	I
M	1.32	F	M	2.03	F	F	1.78	O
M	1.32	F	M	2.31	F	F	1.80	I
M	1.40	F	M	2.36	F	F	1.88	I
M	1.42	I	M	2.46	F	F	2.16	F
M	1.42	F	M	3.25	O	F	2.26	F
M	1.47	I	M	3.28	O	F	2.31	F
M	1.47	F	M	3.33	F	F	2.36	F
M	1.50	I	M	3.56	F	F	2.39	F
M	1.52	I	M	3.58	F	F	2.41	F
M	1.63	I	M	3.66	F	F	2.44	F
M	1.65	O	M	3.68	O	F	2.56	O
M	1.65	O	M	3.71	F	F	2.67	F
M	1.65	I	M	3.89	F	F	2.72	I
M	1.65	F	M	3.89	F	F	2.79	F
M	1.68	F	M	1.24	I	F	2.79	F
M	1.70	I	M	1.30	I	I	2.84	F
M	1.70	I	M	1.45	I	I		
M	1.73	O	M	1.45	O	O		
M	1.78	F	M	1.55	I	I		
M	1.78	O	M	1.55	I	I		
M	1.80	F	M	1.60	I	I		
M	1.85	F	M	1.60	I	I		
M	1.85	F	M	1.65	F	F		
M	1.93	I	M	1.65	F	F		
M	1.93	F	M					
M	1.98	I	M					

Exemple Alligators (2)

The CATMOD Procedure

Maximum Likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	9.84	0.0073
sex	2	2.71	0.2574
length	2	10.28	0.0059
length*sex	2	2.57	0.2767
Likelihood Ratio	94	77.64	0.8890

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	92.270 ^a	.000	0	.
LENGTH	110.319	18.049	2	.000
SEX	95.732	3.461	2	.177

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Exemple Alligators (2)

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	104.563	14.247	2	.001
LENGTH	106.681	16.365	2	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Parameter Estimates

CHOICE	B	Std. Error	Wald	df	Sig.
F Intercept	.998	1.176	.721	1	.396
LENGTH	.085	.489	.030	1	.862
I Intercept	5.181	1.746	8.807	1	.003
LENGTH	-2.388	.921	6.718	1	.010

Exemple Alligators (2)

$$\text{Prob(F)} = \frac{e^{.998+.085\text{Length}}}{1 + e^{.998+.085\text{Length}} + e^{5.181-2.388\text{Length}}}$$

$$\text{Prob(I)} = \frac{e^{5.181-2.388\text{Length}}}{1 + e^{.998+.085\text{Length}} + e^{5.181-2.388\text{Length}}}$$

$$\text{Prob(O)} = \frac{1}{1 + e^{.998+.085\text{Length}} + e^{5.181-2.388\text{Length}}}$$

Exemple Alligators (2)

