1.1 Régression logistique

Bibliographie :

Howell, D.C., Méthodes Statistiques en Sciences Humaines, De Boeck, Paris Bruxelles, 1998.
Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.

1.1.1 La régression logistique

La régression logistique peut être vue comme une extension de la régression linéaire au cas où la variable dépendante est dichotomique. Plus précisément, sur un échantillon de n individus statistiques, on a observé :

- p variables numériques ou dichotomiques X_1, X_2, ..., X_p (variables indépendantes ou explicatives)

- une variable dichotomique Y (variable dépendante, ou "à expliquer").

Dans le cas le plus simple, on cherche à expliquer une variable dichotomique Y par une variable numérique X. On dispose donc d'un tableau de données sous la forme :

	s1	s2		sn
Y	1	0	...	0
X	x1	x2		xn

Exemple : On considère un échantillon de 30 sujets pour lesquels on a relevé :

- d'une part le niveau des revenus (variable numérique)

- d'autre part la possession ou non d'un nouvel équipement électro-ménager.

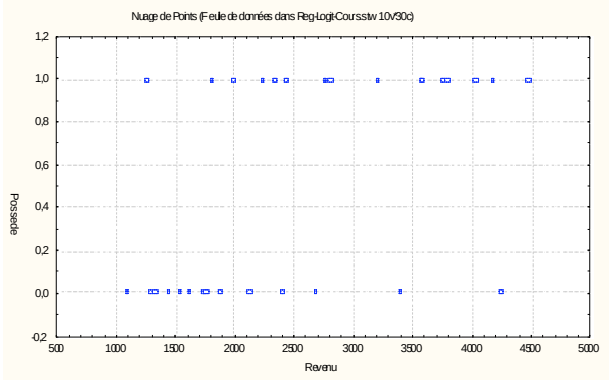
On a obtenu les données suivantes :

Revenu	1085	1304	1331	1434	1541	1612	1729	1759	1863	2121	2395	2681	3390	4237	1241
Possède	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Revenu	1798	1997	2234	2346	2436	2753	2813	3204	3564	3592	3762	3799	4037	4168	4484
Possède	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

1.1.1.1 Principe de la méthode

Ces données peuvent être représentées à l'aide d'un nuage de points, qui a l'allure suivante :



On cherche un modèle permettant d'estimer Y ("Possède") connaissant X ("Revenu"). Plutôt que de rechercher un modèle mathématique donnant pour une valeur donnée X exactement la valeur 0 ou la valeur 1, il peut sembler pertinent de rechercher un modèle produisant des valeurs comprises entre 0 et 1 qui seront interprétées comme des probabilités. Par exemple :

$\hat{Y} = 0{,}1$ signifie que : il y a 10% de chances que Y=1

Cependant, la droite de régression de la variable Y par rapport à la variable X ne constitue pas un bon modèle car les valeurs estimées ne seront pas limitées à 0 et 1.

Pour passer d'une variable prenant ses valeurs dans [0, 1] à une variable prenant ses valeurs dans [0, +∞[, on introduit le rapport de chances ou cote :

$$p_1 = \frac{P(Y=1)}{1-P(Y=1)}$$

Ainsi, si P(Y=1)=0,9, le rapport de chances vaut p1 = 0,9/0,1=9 : on a 9 fois plus de chances d'observer Y=1 que Y=0.
De même, si P(Y=1)=0,2, le rapport de chances vaut p1 = 0,2/0,8=1/4 : on a 4 fois plus de chances d'observer Y=0 que Y=1.

Pour passer d'une quantité (le rapport de chances) variant dans [0, +∞[à une quantité prenant n'importe quelle valeur réelle, on applique une nouvelle transformation, en prenant le logarithme népérien du rapport. On obtient ainsi la transformation logit :

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

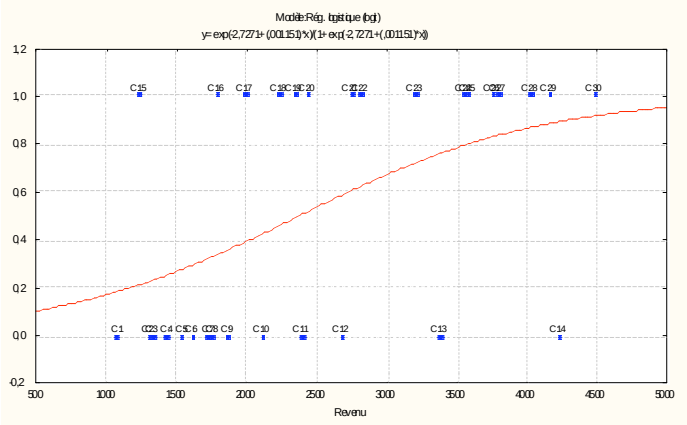
Ainsi,

- si P = 0,9, logit(P) = ln 9 = 2,1972
- si P = 0,5, logit(P) = ln 1 = 0
- si P = 0,2, logit(P) = ln(1/4) = -1,3863.

A partir d'une "valeur logit" y, on peut facilement revenir à la probabilité P correspondante en appliquant la transformation :

$$P = \frac{e^y}{1+e^y}$$

On ajuste alors logit(P) par une fonction affine, ce qui revient à déterminer une "sigmoïde" qui passe au mieux par les points expérimentaux :



L'équation correspondant à cet ajustement est :

$$\text{logit}(Y) = -2{,}7271 + 0{,}001151\, X$$

Exemple d'utilisation de cette équation : à partir de quel revenu a-t-on 90% de chances de tirer un sujet possédant l'équipement envisagé ?

P = 0,9 correspond à P/(1-P) = 0,9/0,1 = 9 d'où logit(P) = 2,1972.
Or : 2,1972 = -2,7271 + 0,001151 X donne X = (2,1972 + 2,7271)/0,001151, c'est-à-dire : X=4278.

Remarque : Cette équation n'est pas obtenue par une "simple" régression linéaire, mais par des méthodes itératives. D'une part, il n'est pas envisageable de faire les calculs manuellement, d'autre part, il faudra, dans certains cas, "aider" les logiciels en indiquant des valeurs initiales plausibles pour les coefficients.

1.1.1.2 Aides à l'interprétation. Evaluation de la qualité du modèle obtenu.

La qualité du modèle peut être évaluée en comparant les résultats obtenus avec ceux du modèle "constant" qui attribuerait la probabilité 14/30 à la valeur 0 et 16/30 à la valeur 1. Une fonction de vraisemblance est évaluée dans les deux cas, et la différence des deux fonctions suit une loi du khi-2 à 1 degré de liberté lorsqu'il n'y a qu'une seule variable indépendante.

Sur notre exemple, on obtient :
Chi-deux = 7,636181 ; dl = 1 ; p = ,0057242
Le revenu est donc un prédicteur significatif de la variable Y.

Une autre aide à l'interprétation courante est le rapport de cotes ou odds-ratio (OR). En particulier, la contribution de la variable X à la variation de Y est calculée par :

$$\text{OR} = \exp(\text{Coefficient de X dans le modèle})$$

Ainsi, sur notre exemple, l'odds-ratio correspondant au coefficient 0,001151 est : $e^{0{,}001151}$=1,0012. Autrement dit, une augmentation du revenu de 1 unité se traduit par une multiplication de la probabilité par 1,0012.

D'une manière générale, l'odds-ratio est défini comme le rapport de deux rapports de chances. Ainsi, l'odds-ratio relatif à l'étendue des valeurs observées est défini de la manière suivante :

- On calcule le rapport de chances relatif à la plus grande valeur observée du revenu :

Pour X = 4484, P_1=0,919325 et $\dfrac{P_1}{1-P_1} = 11{,}3954$

- On calcule le rapport de chances relatif à la plus petite valeur observée du revenu :

Pour X = 1085, P_2=0,185658 et $\dfrac{P_2}{1-P_2} = 0{,}2280$

- L'odds-ratio est obtenu comme quotient des deux rapports précédents :

$$\text{OR} = \frac{\dfrac{P_1}{1-P_1}}{\dfrac{P_2}{1-P_2}} = \frac{11{,}3954}{0{,}2280} = 49{,}98$$

On évalue également un Odds-ratio comparant valeurs observées et valeurs prévues. Pour cela, on définit deux classes dans les valeurs prévues : celles inférieures à 0,5 et celles supérieures à 0,5 et on forme le tableau de contingence croisant les valeurs observées (0 ou 1) avec les classes ainsi définies. Sur notre exemple, on obtient :

Obs	Prév. < 0,5	Prév. > 0,5
0	10	4
1	5	11

Le rapport est alors obtenu en formant le rapport ad/bc (produit des effectifs des cases d'accord divisé par le produit des effectifs des cases de désaccord.
On obtient ainsi :

$$OR = \frac{10 \times 11}{5 \times 4} = 5{,}50$$

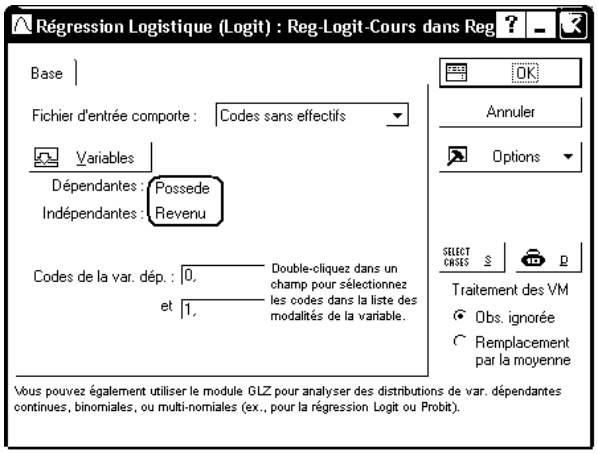
Signification approximative : si la valeur prévue est supérieure à 0,5, on a 5,5 fois plus de chances d'observer Y=1 que Y=0.

1.1.2 La régression logistique avec Statistica

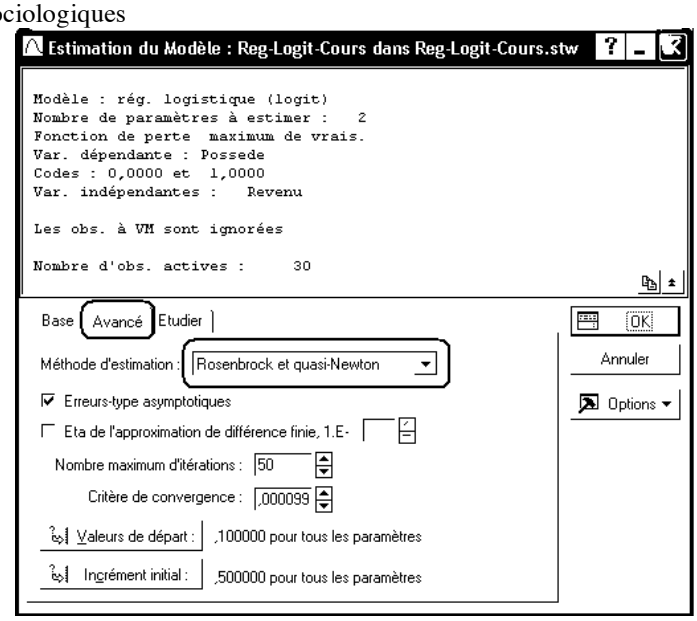
1.1.2.1 Traitement de l'exemple précédent avec Statistica

Le fichier de données se trouve dans le classeur Reg-Logit-Cours.stw

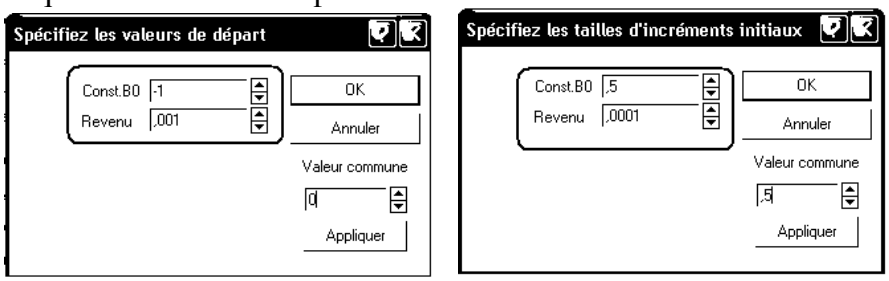
On peut utiliser le menu Statistiques, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit: On indique la variable dépendante et les variables indépendantes :



Cliquez ensuite sur le bouton OK.
Si on se borne à accepter les paramétrages par défaut dans le dialogue suivant, la méthode itérative utilisée par Statistica n'aboutit pas. Il faut donc, soit utiliser un algorithme autre que "Quasi-Newton", proposé par défaut, soit donner des estimations pertinentes des coefficients b0 et b1. Par exemple, on peut spécifier "Rosenbrock et quasi-Newton" comme méthode d'estimation :



On peut aussi indiquer les valeurs de départ et l'incrément initial comme suit :



On obtient comme résultats :

	Modèle: Rég. logistique (logit) Nbre de 0 : 14 1 : 16 Var dép. : Possede Perte : Max vraisemblance (MC-er. posit. à Perte finale= 16,909608827 Chi?(1)=7,6362 p=,00572	
N=30	Const.B0	Revenu
Estimat.	-2,7271	0,0012
Erreur-type	1,2202	0,0005
t(28)	-2,2350	2,4050
niveau p	0,0336	0,0230
-95%CL	-5,2265	0,0002
+95%CL	-0,2277	0,0021
Chi? de Wald	4,9954	5,7838
niveau p	0,0254	0,0162
Odds ratio (unité)	0,0654	1,0012
-95%CL	0,0054	1,0002
+95%CL	0,7963	1,0021
Odds r. (étendue)		49,9828
-95%CL		1,7859
+95%CL		1398,8800

Activer également l'onglet Résidus et étudiez comment est formé le tableau "Classifications d'obs et odds-ratio" à partir du tableau

	Classification d'obs. (Reg-Logit-Cours) Odds ratio : 5,5000 Pourc. corrigé : 70,00%		
Observée	Prév. 0,000000	Prév. 1,000000	% Corrigé
0,000000	10	4	71,42857
1,000000	5	11	68,75000

	Modèle : (Reg-Logit-Cours dans Reg-Logit-Cours.stw) Var. Dép. : Possede		
	Observée	Prév.	Résidus
1	0,000000	0,185658	-0,185658
2	0,000000	0,226805	-0,226805
3	0,000000	0,232300	-0,232300
4	0,000000	0,254106	-0,254106
5	0,000000	0,278143	-0,278143

Le graphique donné dans le paragraphe précédent pourra être obtenu à l'aide du bouton "Fonction 2D ajustée et valeurs observées" de l'onglet "Avancé".

1.1.2.2 Exemple comportant plus d'un prédicteur

Source : Howell. p. 633, ex. 15.31 a 15.33

La feuille de données Harass contient des données légèrement modifiées relatives à 343 cas créés pour répliquer les résultats d'une étude sur le harcèlement sexuel (Brooke et Perot 1991). Les variables sont :
- l'âge
- l'état-civil (1 = marié(e), 2 = célibataire) (NB étonnant, n'est-ce pas l'inverse? cf données)
- l'idéologie féministe
- la fréquence du comportement
- le caractère agressif du comportement
- le fait qu'il ait été ou non signale (0 = non, 1 = oui).

1) Utiliser un programme de régression logistique et examiner la probabilité qu'un sujet signale un cas de harcèlement sexuel sur la base des VI.

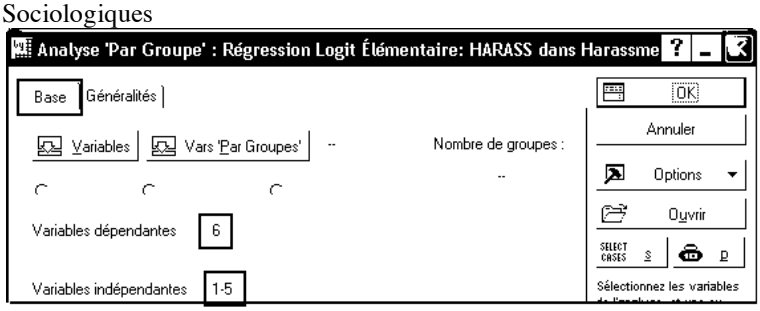
2) Même question, mais en n'utilisant que le prédicteur dichotomique relatif à l'état civil. Faire une table de contingence, calculer les rapports de chances et comparer ces résultats à ceux de la régression logistique. (résultats son significatifs, mais cela importe peu, selon Howell).

3) Apparemment, la fréquence du comportement n'est pas liée à la probabilité de voir la victime signaler le cas de harcèlement. Peut-on en imaginer les raisons ?

Ouvrez le classeur Harassment.stw.

Un premier résultat peut être obtenu à l'aide du menu : Statistiques, Analyses par groupes, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit élémentaire.

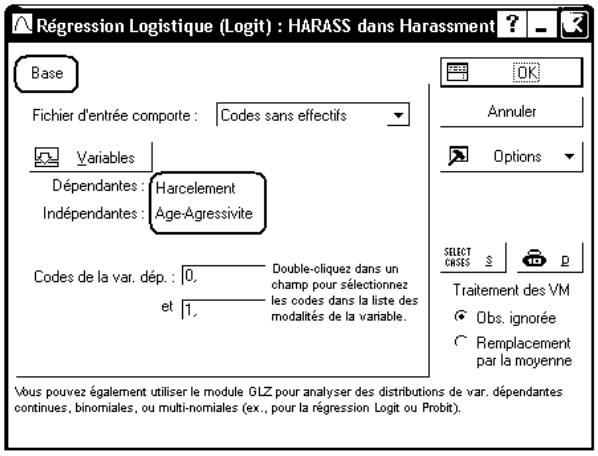
Indiquez "Harcelement" comme variable dépendante et les 5 autres variables comme variables idépendantes.



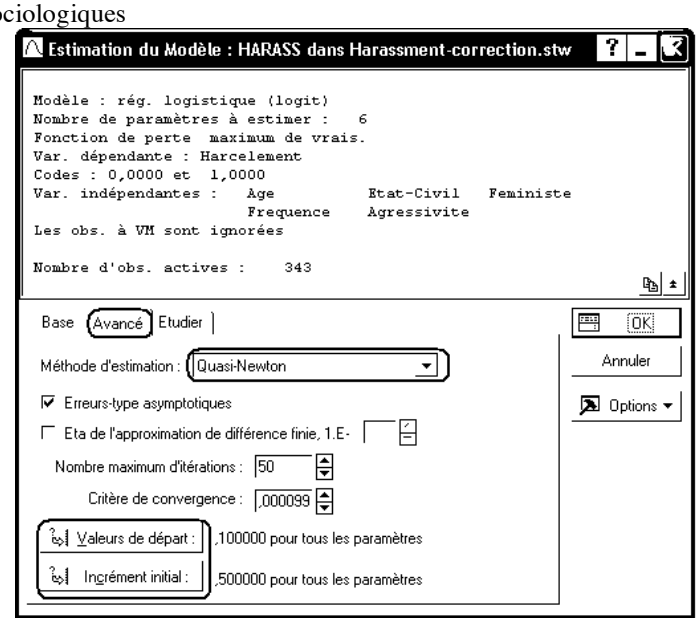
Lorsqu'on laisse les options par défaut (minimum de résultats), on obtient la feuille de résultats suivante :

	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi?(5)=35,442 p=,00000					
N=343	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Estimat.	-1,7317	-0,0137	-0,0723	0,0070	-0,0464	0,4878
Erreur-type	1,4296	0,0129	0,2338	0,0146	0,1525	0,0949
t(337)	-1,2113	-1,0614	-0,3091	0,4771	-0,3043	5,1409
niveau p	0,2266	0,2893	0,7575	0,6336	0,7611	0,0000
-95%CL	-4,5439	-0,0391	-0,5321	-0,0218	-0,3464	0,3011
+95%CL	1,0804	0,0117	0,3876	0,0358	0,2536	0,6744
Chi? de Wald	1,4672	1,1265	0,0955	0,2277	0,0926	26,4292
niveau p	0,2258	0,2885	0,7573	0,6333	0,7609	0,0000
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287
-95%CL	0,0106	0,9617	0,5874	0,9784	0,7072	1,3514
+95%CL	2,9460	1,0118	1,4734	1,0364	1,2887	1,9629
Odds r. (étendue)		0,4644	0,9303	1,3985	0,8306	80,6394
-95%CL		0,1121	0,5874	0,3509	0,2501	15,0336
+95%CL		1,9244	1,4734	5,5731	2,7579	432,5462

On peut aussi utiliser le menu Statistiques, Modèles linéaires/non-linéaires avancés, Estimation non linéaire, Régression Logit: On indique de la même façon la variable dépendante et les variables indépendantes :



On peut ensuite choisir un algorithme d'estimation et éventuellement indiquer manuellement les valeurs initiales des coefficients bi, ce qui est souvent utile, si les plages de variations des VI sont très différentes de l'intervalle [0, 1] (et n'est pas prévu par le menu précédent).



Le tableau de résultats produit par la méthode précédente est alors accessible par le bouton "Synthèse : paramètres et erreurs-types" du dialogue des résultats.

L'équation de la courbe de régression est :

$$\text{logit}\, P = -1{,}7317 - 0{,}013698\,\text{Age} - 0{,}072251\,\text{EtatCivil} + 0{,}0069870\,\text{Feministe} - 0{,}046408\,\text{Frequence} + 0{,}4878\,\text{Agressivite}$$

Le khi-2 correspondant au modèle vaut 35,442, et il est significatif au seuil de 1%. En revanche, seule la variable Agressivite semble avoir un rôle explicatif supérieur à celui que le hasard est susceptible de produire.

Les odds-ratio unitaires correspondant aux différentes variables sont :

	Modèle: Rég. logistique (logit) Nbre de 0 : 174 1 : 169 (HARASS) Var dép. : Harcelement Perte : Max vraisemblance (MC-er. posit. à Perte finale= 219,99193498 Chi?(5)=35,442 p=,00000					
N=343	Const.B0	Age	Etat-Civil	Feministe	Frequence	Agressivite
Odds ratio (unité)	0,1770	0,9864	0,9303	1,0070	0,9547	1,6287

On voit que seules les variables Feministe et Agressivite possèdent des odds-ratio unitaires supérieurs à 1 et que seul celui de Agressivite est nettement différent de l'unité.

On peut également afficher le tableau des valeurs observées et des valeurs prévues de la variable dépendante :

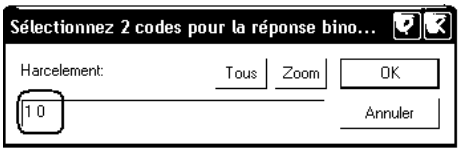
	Modèle : (HARASS dans Harassment-correction.stw) Var. Dép. : Harcelement		
	Observée	Prév.	Résidus
1	0,0000	0,6431	-0,6431
2	0,0000	0,7312	-0,7312
3	1,0000	0,8696	0,1304
4	1,0000	0,3080	0,6920

Sous l'onglet Résidus, on peut obtenir le calcul de l'odds-ratio pour le modèle :

	Classification d'obs. (HARASS) Odds ratio : 2,1051 Pourc. corrigé : 59,18%		
Observée	Prév. 0,000000	Prév. 1,000000	% Corrigé
0,000000	111	63	63,79310
1,000000	77	92	54,43787

On peut également utiliser le menu : Statistiques, Modèles linéaires/non-linéaires avancés, Modèles linéaires/non linéaires généralisés, puis l'item Modèle logit dans l'onglet Base ou les items : Régression simple (ou multiple), Distribution: Binomiale et Fonction de liaison : logit de l'onglet Avancé.

Lorsqu'on indique les variables et leur rôle, il est important de préciser que c'est le code "1" de la variable Harcelement qui doit être assimilé à la modalité "succès" de la variable binomiale, faute de quoi les résultats seraient inversés :



On retrouve ainsi les résultats obtenus par les deux autres méthodes, mais avec une présentation différente. On peut également obtenir des résultats supplémentaires, tels que l'évolution des valeurs des coefficients à chaque itération de l'algorithme :

	Harcelement - Historique itérations (HARASS) Distribution : BINOMIALE Fonction de Liaison : LOGIT					
Effet	Niveau Effet	Colonne	Itérat. 0	Itérat. 1	Itérat. 2	Itérat. 3
Ord.Orig		1	0,000	-1,556	-1,727	-1,732
Age		2	0,000	-0,012	-0,014	-0,014
Etat-Civil		3	0,000	-0,066	-0,072	-0,072
Feministe		4	0,000	0,006	0,007	0,007
Frequence		5	0,000	-0,044	-0,046	-0,046
Agressivite		6	0,000	0,438	0,486	0,488
Vraisembl.			-237,749	-220,156	-219,992	-219,992

On peut également noter que l'on obtient des résultats légèrement différents lorsque l'on indique "Etat-Civil" comme variable catégorielle.

1.1.3 Un exemple de régression logistique issu d'un article.

Réf. : Factors Influencing Adolescents Engagement in Risky Internet Behavior, ALBERT KIENFIE LIAU, Ph.D., ANGELINE KHOO, Ph.D., and PENG HWAANG, Ph.D., CYBERPSYCHOLOGY & BEHAVIOR, Volume 8, Number 6, 2005, pp 513-520.

Dans l'article cité supra les auteurs se sont intéressés aux facteurs liés à la prise de risques dans le comportement sur Internet pour des adolescents de Singapour. Ils identifient notamment comme conduite à risques le fait de rencontrer physiquement une personne qu'ils ont d'abord connu "online".

Dans les résultats de leur étude, les auteurs indiquent notamment :

1045 (93.0% of the total sample) adolescents reported having used the Internet, and 827 (73.6%) adolescents reported having chatted on the Internet. The study focused on this group of 827 adolescents who have experienced chatting on the Internet. These adolescents have a mean age = 14.42 (SD = 1.33) and are 51.4% girls. (...)
A total of 169 adolescents (16.2% of Internet users, or 20.4% of those who chat) reported having met someone in real life that they first encountered online.

A series of multiple logistic regression analyses was used to examine the factors that influence adolescents' engagement in risky internet behavior, in particular, meeting in person with someone encountered online. Odds ratios (OR) were calculated to approximate relative risk and are presented with 99% confidence intervals. Age was a significant predictor of the risky behavior (OR = 1.26, 99% CI (1.06, 1.48), $p < 0.0001$) but gender was not a significant predictor; 80 out of the 169 (47.3%) adolescents were girls. For ease of interpretation, the frequency of use of the Internet variable was dichotomized so that 1 = "at least once a day" and 0 = "less than once a day." Controlling for age, frequency of use of the Internet was a significant predictor of the risky behavior (OR = 1.68, 99% CI (1.07, 2.65), $p < 0.01$). Parents' educational background and whether parents lived together were not significant predictors of the risky behavior. All subsequent analyses include age and frequency of use as covariates in order to control for the influence of these factors. The following factors were examined as predictors of the risky behavior: frequency of chatting and gaming behavior, parental supervision, communication with parents, type of personal information given out, amount of inappropriate messages received, whether inappropriate websites have been visited, and type of internet advice heard. Significant and marginally significant predictors of the risky behavior are reported in Table 2.

TABLE 2. SIGNIFICANT AND MARGINALLY SIGNIFICANT PREDICTORS OF THE RISKY INTERNET BEHAVIOR—MEETING IN PERSON SOMEONE ENCOUNTERED ONLINE

Predictor	*OR*	*99% CI*
Frequency of Internet activities	3.13**	1.75, 5.55
Frequency of chatting	1.77*	1.07, 2.91
Frequency of gaming		
Parental supervision		
Rules for Internet use		
Not allowed to meet in person someone encountered online	0.49**	0.30, 0.81
Not allowed to talk to strangers in chatrooms	0.46*	0.23, 0.93
Not allowed to give out personal information	0.62†	0.39, 1.01
People usually at home when arrive from school	1.56†	1.06, 1.48
Communication with parents		
Tell parents about receiving pornographic junk mail	0.49†	0.22, 1.06
Giving out personal information		
Phone number	2.17*	1.15, 4.09
Photograph	2.68*	1.16, 6.18
Favorite band, music	1.67*	1.03, 2.90
Receiving inappropriate message		
Met someone on the Internet who asked for personal information	4.16**	2.42, 6.67
Sent pornography from someone met only on the Internet	1.80†	0.97, 3.34
Received unwanted sexual comments on the Internet	2.59**	1.58, 4.23
Received pornographic junk mail in e-mail or Instant Messaging	1.90**	1.19, 3.04
Visiting Inappropriate websites		
Accidentally ended up in a pornographic website	1.68*	1.04, 2.73
Purposely visited a pornographic website	2.39**	1.33, 4.28
Accidentally ended up in a website with violent/gruesome images	1.60*	1.01, 2.54
Accidentally ended up in a hate website	1.44†	0.90, 2.33
Heard of the following Internet safety advice		
Never arrange to meet anyone	0.55*	0.33, 0.90
Do not download anything	1.88*	1.06, 3.17

**$p < 0.0001$.
*$p < 0.01$.

$+p < 0.05$.