Séance 8 : Régression Logistique

Sommaire

La régression logistique traite :

- Une relation d'explication linéaire d'une transformation de Y : G(Y) = a+b.x.
  - La fonction de lien (Link =) est un Logit : Log[Y / (1-Y)]
  - La fonction est logistique : Y = exp(a+b.x) / (1+ exp(a+b.x)
- Avec une variable à expliquer (Y)
  - Nominale (0/1) : Régression logistique binaire simple (à une variable explicative) ou multiple.
  - Ordinale : régression logistique ordinale
  - Multinomiale (choix de 1 option parmi n) : MNL

La régression logistique est basée sur une transformation de la VAE par un logit. On peut aussi mettre dans cette catégorie de régression binaire

- la régression Probit ou Normit (basée sur une distribution Normale)
- le modèle Log-Log : g(Y) = log(-log(1-Y))
- le modèle Tobit (avec un emboitement d'une logistique binaire et d'un MNL)
- les régressions logistiques « nichées » : pour lesquelles le choix d'une option est emboité dans un choix précédent.

Dans sa version simple, la régression logistique est soumise à une hypothèse forte : l'indépendance des alternatives non pertinentes (IIA) qu'il faut étudier.

- selon cette hypothèse IIA le choix entre deux options ne dépend pas de la présence/absence d'autres options :
- le choix entre un taxi et un bus ne dépend pas du fait que le bus soit disponible en rouge ou en bleu.

L'estimation des coefficients se fait par la maximisation du Logarithme de la vraisemblance.

- La vraisemblance est la probabilité d'observer l'échantillon compte tenu du modèle et de ses paramètres
- Exemple : si binomial avec une probabilité « p » d'avoir un résultat « pair » alors la probabilité d'observer un échantillon avec 3 pair et 2 impair est p.p.p.(1-p).(1-p) soit p^3.(1-p)^2
- Pour rendre plus facilement manipulable la vraisemblance on passe d'un produit à une somme en passant par les Log et en prenant le négatif pour minimiser.

La régression est dite

- Ordinaire si l'échantillon est tiré au hasard.
- Conditionnelle si les « Y » sont connus au moment de la constitution de l'échantillon (échantillons appariés : on recherche des individus ayant les même caractéristiques que ceux pour qui l'évènement est survenu). C'est le cas si l'échantillon est stratifié.

Proc LOGISTIC : Régression logistique

Principe : Explication d'un choix ou de la survenue d'un évènement.

Nature des variables et de la distance :

- Y = variable discrète correspondant à un choix
- X : variables nominales ou quantitatives

Remarque :

Effectifs minima d'environ 10 pour que les tests de validation soient valables

Exemple simple de Hosmer et Lemeshow sur le risque d'accident cardio-vasculaire selon l'âge

```
*************************************************;
* logistique binaire avec variables mixtes;
*************************************************;
ods graphics on ;
 Proc logistic data=in_200
        plots (only) = (effect oddsratio roc) ;
        class sexe  (ref='F')    / param = ref;
        model achat (event="1") = catotal  sexe /
                LINK=logit  /* LINK=logit demande une régression logistique*/
                RSQUARE       /* RSQUARE demande le R2 */
                LACKFIT
                EXPB            /* EXPB expression des exponentielles des coefficients*/
                outroc= roc;      /* courbe ROC*/
        oddsratio catotal;
        oddsratio sexe ;
        output out=data_out    predprobs=I p=prob xbeta=logit resdev=resdev;
run ;

**** recherche des points à problème ******************;
*************************************************************;
title3 "points à problème" ;
proc print data=data_out;
        where resdev>2 or resdev<-2 ;
run;
```

Décisions à prendre :

- **(0) Choisir la modalité « 1 »** :
  - Il s'agit de prévoir au mieux la modalité « 1 » de la variable Y

- **(1) Signification globale du modèle** : soit LL(cste) = Log-vraisemblance (du modèle Cste)

  - Valeurs de la qualité de l'ajustement lié à la vraisemblance. **Le plus petit, le meilleur est l'ajustement.**
    - On retrouve -2Log L (-2 log de la vraisemblance).
    - AIC (Critère d'information d'Akaike) pénalise du nombre de paramètres
    - SC (critère de Schwartz) prend aussi en compte les effectifs

  - **Test de la qualité d'ajustement (Goodness of fit) de Hosmer & Lemshow :** test du chi² sur les fréquences observées et théoriques (pour une variable continue, le découpage est fait en déciles avec ddl = k-2).

- **Les « R^2 »** sont calculés à partir de ratios de LL(Cste) et de LL(Cste, X)
  - R2 de Cox & Snell
  - Pseudo R2 (McFadden)
  - R2 ajusté de Nagelkerke (par rapport au R^2 maximum possible)

- **Les « bien classés »** : faire un tableau «croisant « prévu » et « réel ». On mesure ainsi les effectifs totaux (N), concordants (nc), discordants (nd) et le nombre total de réponses différentes (t). Ex-aequo =' % lié'
  - Goodman-Kruskal Gamma Gamma = (nc - nd) / (nc + nd) [-1, +1], le plus simple mais sur-estime la relation s'il y a beaucoup d'ex-aequo.
  - *Interprétation* (uniquement pour Gamma): Si = 0.43, connaître la variable explicative réduit de 43% l'erreur de prévision sur Y.
    - *<0.2 faible / 0.4 modérée / 0.6 forte / 0.8 Très forte*
  - D de Somer = (nc - nd) / t prend en compte les ex-aequo.
  - Tau-a = (nc - nd) / .5N(N-1) corrige le D de l'importance globale des ex-aequo par rapport au total des réponses.
  - C [0.5, 1] = aire sous la courbe ROC (voir ci-dessous)
  - http://salises.mona.uwi.edu/sem2_09_10/SALI6031/Ordinal%20Measures%20of%20Association.htm

- **La courbe ROC**
  - Sensibilité : capacité à prévoir « 1 » pour les « 1 » réels
  - Spécificité : capacité à prévoir « 0 » pour les « 0 » réels.
  - 1-Spécificité : risque de prévoir « 1 » pour les « 0 » réels.
  - Courbe ROC : Y= Sensibilité / X = 1- Spécificité)
    - Il faut trouver un compromis entre la maximisation de ces deux indicateurs
- La courbe Lift (levier)
  - Lien entre un seuil de sélection et la sensibilité (même ordonéne que la courbe ROC).

- **(2) Signification du coefficient d'une variable**
  - **Le test LRT (Likelihood ratio test)** (à préférer à Wald)**:**
    - -2LL(Cste)-(-2LL(Cste, X)
    - H0 = le coefficient est nul

  - **Le test de Wald** doit être signification (H0 : le coefficient est nul)
    - C'est le rapport du carré du coefficient sur sa variance
    - C'est la racine carrée du chi² donné par SAS

  - **Le rapport de côtes (odds ratio)**
    - En considérant toutes les autres variables inchangées
    - La côte est la probabilité qu'un évènement se réalise sur la proba qu'il ne se réalise pas (pour un pari : '15 contre 1')
    - Le rapport de côte calculé est l'exponentielle du coefficient estimé. Il indique de combien va varier le rapport de côte si l'on modifie X d'une unité.
    - Un rapport de côte > 1 indique que la probabilité de survenue de l'évènement est plus élevée.

Exemple

Table de SEXE par ACHAT				
		ACHAT		
		0	1	Total
SEXE				
F	Fréquence	106	94	200
	Pctage en ligne	53.00	47.00	
M	Fréquence	62	138	200
	Pctage en ligne	31.00	69.00	
Total	Fréquence	168	232	400

	0	1	Côte (odd)
Femme	106	94	0,887
Homme	62	138	2,226
Rapport de côtes (odds ratio)			**0,398**

- **(3) Détection des points individuels à problème**
  - A partir de l'analyse des résidus standardisés de Pearson (écart prévu réel divisé par l'écart-type de la valeur prévue (racine(p*q))
    - A comparer à 2 en valeur absolue
  - La déviance = -2LL
  -
- **(4) Possibilité de sélectionner automatiquement les variables**
- `Dans model : / SELECTION=FORWARD COVB LACKFIT CLODDS CLPARM;`

Exemple commenté : Achat en (t+1) à partir du sexe et du chiffre d'affaires de la période précédente

Profil de réponse		
Valeur ordonnée	ACHAT	Fréquence totale
1	0	168
2	1	232

La probabilité modélisée est ACHAT=1.

Informations sur le niveau de classe		
Classe	Valeur	Variables d'expérience
SEXE	F	1
	M	-1

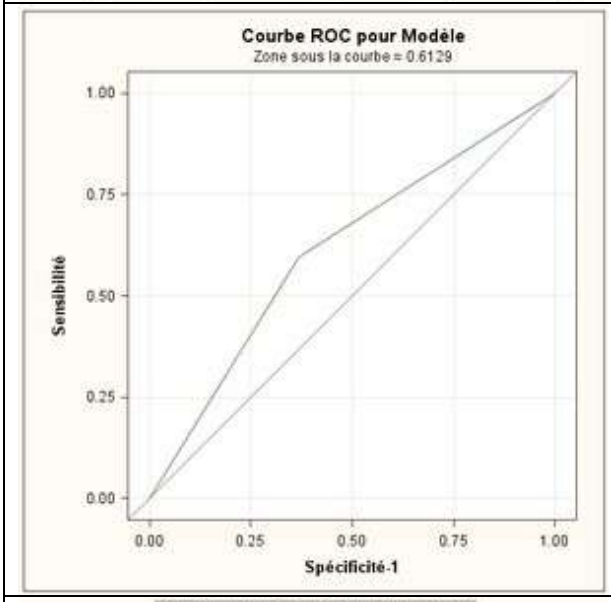
Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et covariables
AIC	546.234	528.179
SC	550.225	536.162
-2 Log L	544.234	524.179

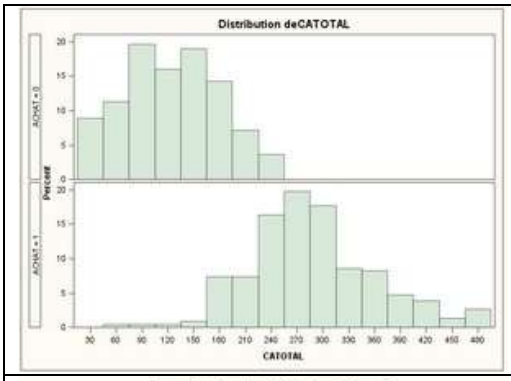
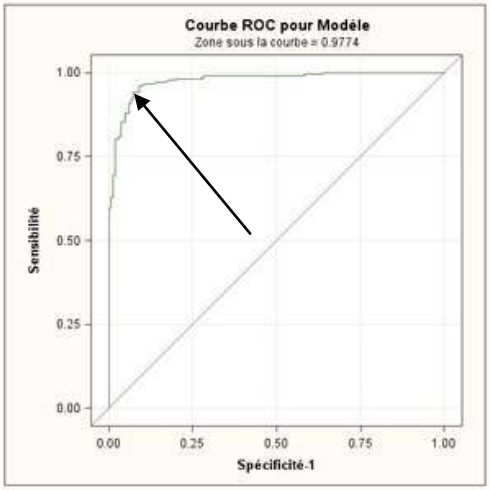
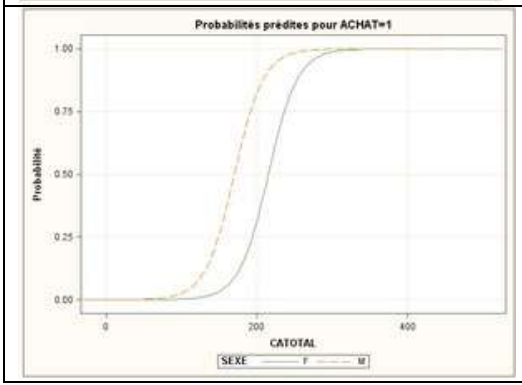
R carré	0.0489	R carré remis à l'échelle max.	0.0658

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapp. de vrais.	20.0549	1	<.0001
Score	19.8686	1	<.0001
Wald	19.4872	1	<.0001

Rappel du profil des réponses
Rappel de la modalité qui est prévue (ici achat =1)
Information sur le codage de la variable : ici codage symétrique (+1/-1) et non pas codage en dummy (0/1)
On retrouve -2Log L (-2 log de la vraisemblance). **Le plus petit, le meilleur est l'ajustement.** Les deux autres pénalisent le modèle par le nombre de variables explicatives **(**AIC, Critère d'information d'Akaike ; SC, critère de Schwartz)
L'apport de la variable explicative est mesuré par l'écart entre le modèle de base (Constante) et le modèle avec variables explicatives. Ici entre 546 et 528.
Le pourcentage de variance expliquée est donc faible 0.0489.
Comme le maximum du R^2 ne peut pas toujours être égal à 1, on peut recalculer le R2 (Nagelkerke) par rapport à son maximum (ici 0.0658)
Le test global (H0 tous les coefficients sont nuls) rejette H0.

Analyse des effets Type 3 Effet / DDL / Khi-2 de Wald / Pr > Khi-2 SEXE / 1 / 19.4872 / <.0001 **Estimations par l'analyse du maximum de vraisemblance** Paramètre / / DDL / Valeur estimée / Erreur type / Khi-2 de Wald / Pr > Khi-2 / Exp(Est) Intercept / / 1 / 0.3399 / 0.1042 / 10.6384 / 0.0011 / 1.405 SEXE / F / 1 / -0.4601 / 0.1042 / 19.4872 / <.0001 / 0.631 **Estimations des rapports de cotes** Effet / Valeur estimée du point / Intervalle de confiance de Wald à 95 % / SEXE F vs M / 0.398 / 0.265 / 0.600 **Association des probabilités prédites et des réponses observées** Pourcentage concordant / 37.5 / D de Somers / 0.226 Pourcentage discordant / 15.0 / Gamma / 0.430 Pourcentage lié / 47.5 / Tau-a / 0.110 Paires / 38976 / c / 0.613 **Intervalle de confiance de Wald pour les rapports de cotes** Libellé / Valeur estimée / Intervalle de confiance à 95 % / SEXE F vs M / 0.398 / 0.265 / 0.600	La variable sexe est significative (probabilité < 5%). Le coefficient estimé est de -0.46 (son exponentielle vaut 0.63). Le rapport de côtes (odds ratio) vaut 0.398 (cf résultat avec le tableau croisé). Par rapport à un Homme, Etre une femme réduit la côte d'un « Achat=1 » de 60,2% (1-0.398) Le tableau de concordance (« bien classés ») donne différents indicateurs qui concluent tous à une relation modérée à faible (force et sens de la relation) C [0.5, 1] (aire sous la courbe ROC)
	La courbe ROC donne une valeur de C = 0.61 (déjà présente fans le tableau précédent) Elle est aussi considérée comme faible.
Partition pour les tests de Hosmer et de Lemeshow Groupe / Total / ACHAT = 1 Observé / ACHAT = 1 Attendu / ACHAT = 0 Observé / ACHAT = 0 Attendu 1 / 200 / 94 / 94.00 / 106 / 106.00 2 / 200 / 138 / 137.99 / 62 / 62.01 **Test d'adéquation de Hosmer et de Lemeshow** Khi-2 / DDL / Pr > Khi-2 0.0000 / 0 / **Table de classification** Niveau de proba. / Correct Evénement / Correct Non-événement / Incorrect Evénement / Incorrect Non-événement / Pourcentages Correct / Sensibilité / Spécificité / Faux POS / Faux NEG 0.100 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.150 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.200 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.250 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.300 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.350 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.400 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.450 / 232 / 0 / 168 / 0 / 58.0 / 100.0 / 0.0 / 42.0 / 0.500 / 138 / 106 / 62 / 94 / 61.0 / 59.5 / 63.1 / 31.0 / 47.0 0.550 / 138 / 106 / 62 / 94 / 61.0 / 59.5 / 63.1 / 31.0 / 47.0 0.600 / 138 / 106 / 62 / 94 / 61.0 / 59.5 / 63.1 / 31.0 / 47.0 0.650 / 138 / 106 / 62 / 94 / 61.0 / 59.5 / 63.1 / 31.0 / 47.0 0.700 / 0 / 168 / 0 / 232 / 42.0 / 0.0 / 100.0 / / 58.0 0.750 / 0 / 168 / 0 / 232 / 42.0 / 0.0 / 100.0 / / 58.0 0.800 / 0 / 168 / 0 / 232 / 42.0 / 0.0 / 100.0 / / 58.0 0.850 / 0 / 168 / 0 / 232 / 42.0 / 0.0 / 100.0 / / 58.0 0.900 / 0 / 168 / 0 / 232 / 42.0 / 0.0 / 100.0 / / 58.0	Le test de Hosmer et Lemeshow (à base de Chi²) conduit à conclure que l'ajustement prévu/réel est bon (H0 = le modèle est bien adapté) La table de classification est peu utile ici (cas d'une variable discrète)

Poursuite en intégrant le chiffre d'affaires

	Histogramme du CA selon l'achat ou non. Possibilité aussi de faire une première analyse avec un test en t
Analyse des effets Type 3 Effet / DDL / Khi-2 de Wald / Pr > Khi-2 CATOTAL / 1 / 72.6380 / <.0001 SEXE / 1 / 22.9055 / <.0001 **Estimations par l'analyse du maximum de vraisemblance** Paramètre / DDL / Valeur estimée / Erreur type / Khi-2 de Wald / Pr > Khi-2 / Exp(Est) Intercept / 1 / -8.8317 / 1.1039 / 64.0080 / <.0001 / 0.000 CATOTAL / 1 / 0.0519 / 0.00609 / 72.6380 / <.0001 / 1.053 SEXE F / 1 / -2.3370 / 0.4883 / 22.9055 / <.0001 / 0.097 **Estimations des rapports de cotes** Effet / Valeur estimée du point / Intervalle de confiance de Wald à 95 % CATOTAL / 1.053 / 1.041 / 1.066 SEXE F vs M / 0.097 / 0.037 / 0.252 **Association des probabilités prédites et des réponses observées** Pourcentage concordant / 97.7 / D de Somers / 0.955 Pourcentage discordant / 2.2 / Gamma / 0.955 Pourcentage lié / 0.0 / Tau-a / 0.466 Paires / 38976 / c / 0.977	Les deux variables sont significatives et le poids du sexe est bien augmenté. La qualité globale est maintenant très bonne.
	La courbe ROC s'est bien améliorée. Elle doit être si possible tout en haut à gauche. Interprétation : trouver le plus grand nombre possible de « 1 » (ordonnes) en minimisant le nombre de faux signaux (0) (abscisses).
	Profil des probabilités des deux courbes (sexe) selon le niveau du CA (abscisses)