La régression logistique

Par Sonia NEJI et Anne-Hélène JIGOREL

Introduction

- La régression logistique s'applique au cas où:
  - Y est qualitative à 2 modalités
  - X_k qualitatives ou quantitatives

- Le plus souvent appliquée à la santé:
  - Identification des facteurs liés à une maladie
  - Recherche des causes de décès ou de survie de patients

Plan

I. Spécification du modèle

II. Interprétation des coefficients

III. Estimations et tests des paramètres

IV. Adéquation du modèle

V. Application

I. Spécification du modèle

Contexte

- Y est une variable binaire
  - 0 en cas de non occurrence de l'évènement.
  - 1 si occurrence.
- Y aléatoire et X_i non aléatoires
- On cherche à expliquer la survenue d'un évènement
- On cherche la probabilité de succès
- On travaille en terme d'espérance

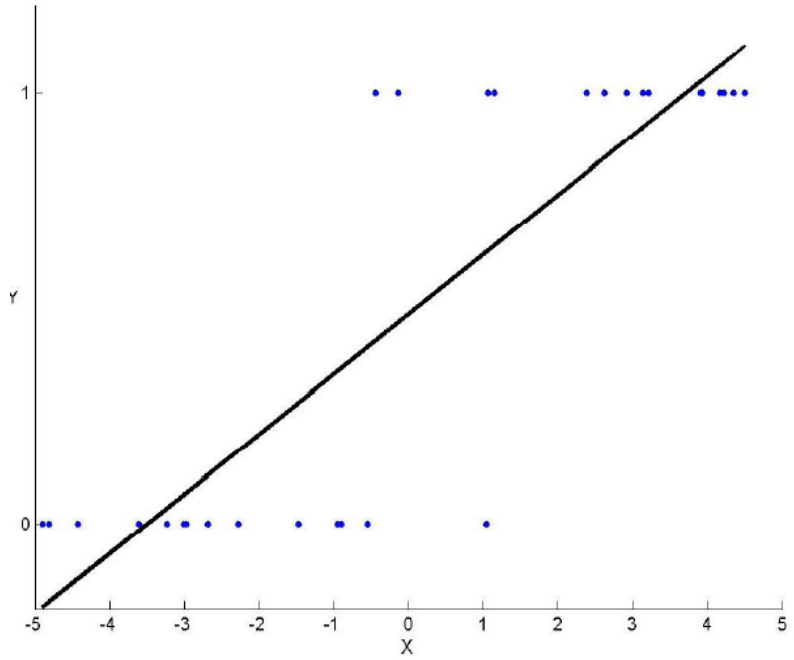
Notations

- On note:
  - $(Y,X_1,X_2,\ldots,X_k)$ les variables de la population dont on extrait un échantillon de n individus i.
  - (y_i,x_i) est le vecteur des réalisations de (Y_i,X_i)
  - K variables explicatives

Contexte

$$Y = f(x_1, x_2, \ldots, x_k)$$

- f ne peut être une fonction linéaire car Y ne prend que deux valeurs:



Contexte

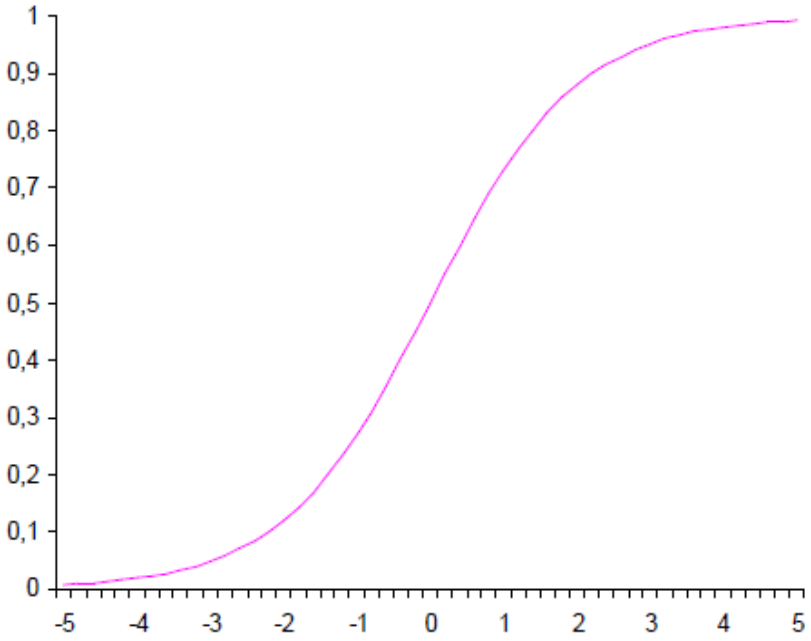
- Afin que l'espérance de Y ne prenne que 2 valeurs, une utilise la fonction logistique :

$$f(x) = \frac{\exp(x)}{1 + \exp(x)} = p$$

  - Ainsi:

    0<f(x)<1

    et E(Y) =0 ou 1



Loi de Y

- Y suit une loi de Bernoulli de paramètre p

- Application de la transformation *logit* permet de travailler sur des valeurs entre [-∞;+∞]:

$$logit(p) = \ln(\frac{p}{1-p})$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ik}$$

II. Interprétation des coefficients

Cas d'une seule variable exogène binaire

- L'Odds (ou « *cote* »)
  - Soit P une probabilité. Son odds est défini par:

$$Odds_P = \frac{P}{1-P}$$

- Par exemple, si un étudiant a 3 chances sur 4 d'être reçu, contre 1 chance sur 4 d'être collé, sa cote est de « 3 contre 1 », soit

$$Odds = \frac{3/4}{1/4} = 3$$

Cas d'une seule variable exogène binaire

- **Odds ratio** (ou « rapport des cotes »)
  - C'est le rapport des cotes des probabilités d'avoir la maladie pour ceux qui ont un symptôme X d'une part et de ceux qui ne l'ont pas d'autre part.

- OR=1, la maladie est **indépendante** du symptôme
- OR>1, la maladie est plus fréquente pour les individus qui **ont** le symptôme.
- OR<1, la maladie est plus fréquente pour les individus qui **n'ont pas** le symptôme.

Cas d'une seule variable exogène binaire

$$RC = \frac{\frac{P(Y_i = 1 \,|X = 1)}{1 - P(Y_i = 1 \,|X = 1)}}{\frac{P(Y_i = 1 \,|X = 0)}{1 - P(Y_i = 1 \,|X = 0)}}$$

$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Cas d'une seule variable exogène binaire

- X=0 : symptôme absent
- X=1 : symptôme présent
- Y=0: la maladie est absente
- Y=1: la maladie est présente

- On a donc:

$$Logit[P(Y_i = 1 \,|X = x)] = \beta_0 + \beta_1 x$$

Cas d'une seule variable exogène binaire

- **Avec l'estimateur de β1**: RC=exp(β_1), permet de comparer les individus qui possèdent le symptôme X avec ceux qui ne le possède pas. Pour cela, on compare le RC à 1.

- **Avec l'estimateur de β0**: On peut calculer

$$P(Y_i = 1 \,|X = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

C'est-à-dire la proportion observée de malades n'ayant pas le symptôme.

Cas d'une seule variable exogène quantitative

- X une variable quantitative *(ex: âge)*
- Y=0: la maladie est absente
- Y=1: la maladie est présente

- On a encore:

$$Logit[P(Y_i = 1 \,|X = x)] = \beta_0 + \beta_1 x$$

Cas d'une seule variable exogène quantitative

- **Avec l'estimateur de β1**: permet d'avoir le l'odds ratio quand X1 augmente d'une unité:

$$RC = \exp(\beta_1)$$

Cas d'une seule variable exogène quantitative

- **Avec l'estimateur de β0**: permet de connaitre la proportion de malades dont la valeur de X est 0.

Attention à l'interprétation de β0 qui n'a pas de sens pour certaines variables X comme l'âge!

Synthèse: Modèle logistique multiple

- L'interprétation est semblable à celle des modèles à une variable explicative.

- Exemple:

$$Logit[P(Maladie_i = 1 \,|Age, fume)] = \beta_0 + \beta_1.Age + \beta_2.fume$$

$$\widehat{\beta_0} = 1{,}3982 \;, \widehat{\beta_1} = 0{,}4118 \; et \; \widehat{\beta_2} = 0{,}6708$$

Synthèse: Modèle logistique multiple

$$Logit[P(Maladie_i = 1 \,|Age, fume)] = \beta_0 + \beta_1.Age + \beta_2.fume$$

$$\widehat{\beta_0} = 1{,}3982 \;, \widehat{\beta_1} = 0{,}4118 \; et \; \widehat{\beta_2} = 0{,}6708$$

- L'interprétation de β0 n'a pas de sens

- RC=exp(β1)=1,5068 >1
- Si l'âge augmente d'une unité, le risque de contracter la maladie **augmente**.

- RC=exp(β2)=1,9558 >1
- Le risque de contracter la maladie est **plus élevé** si l'individu est fumeur.

III. Estimation et test du modèle

Maximum de vraisemblance

- Estimateurs des paramètres sans biais et de faible variance.

- n variables aléatoires Y_i iid qui suivent une loi de $\mathcal{B}(\beta)$.

- La **vraisemblance** d'un n-échantillon $y_1, y_2, \ldots, y_n$ est définie comme la probabilité d'observer cet échantillon.

Maximum de vraisemblance

$$P(Y_i = y_i) = {\beta_i}^{y_i}.(1-\beta_i)^{1-y_i}$$

- Les variables Yi étant indépendantes:

$$\mathrm{L}(\beta,y1\ldots,yn\ ) = \prod_{i=1}^{n} \beta^{y_i}.(1-\beta)^{1-y_i}$$

Maximum de vraisemblance

- Avec $s(\beta_j)$ tel que $s^2(\beta_j)$ soient les variances des estimateurs telles que la matrice de variance covariance soit de la forme :

$$s^2(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} s^2(\hat{\beta}_0) & s^2(\hat{\beta}_0,\hat{\beta}_1) & \cdots & s^2(\hat{\beta}_0,\hat{\beta}_p) \\ s^2(\hat{\beta}_1,\hat{\beta}_0) & s^2(\hat{\beta}_1) & \cdots & s^2(\hat{\beta}_1,\hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ s^2(\hat{\beta}_p,\hat{\beta}_0) & s^2(\hat{\beta}_p,\hat{\beta}_1) & \cdots & s^2(\hat{\beta}_p) \end{pmatrix}.$$

Maximum de vraisemblance

Intervalles de confiance

- Ce test permet de savoir s'il y a une relation entre la variable X_j et Y.

$$IC = \exp[\widehat{\beta_j} \pm u_\alpha . s(\widehat{\beta_j})]$$

- Si 1 ∉ IC ⟶ pas de relation
- Si 1 ∈ IC ⟶ relation entre X_j et Y

Test du rapport de vraisemblance

- Compare 2 modèles emboités:
  - M1: k paramètres
  - M2: p paramètres (p>k)

- Les hypothèses de test sont:

$$\begin{cases} H_0 : On\ choisit\ M1\ (même\ qualité\ prédictive\ que\ le\ 2\ mais\ moins\ de\ paramètres) \\ H_1 : On\ choisit\ M2\ (meilleure\ qualité\ prédictive) \end{cases}$$

- La statistique de test est:
  - (-2.ln(vraisemblance au maximum de M1)] - (-2.ln(vraisemblance au maximum deM2)]
- Elle suit une loi du **Khi-deux** à **p-k** degrés de libertés.

Test de significativité globale

- Les variables explicatives influencent-elles simultanément le risque de survenue de l'événement?

- On va effectuer un test du rapport de vraisemblance…

Test de significativité globale

- M1: Modèle **sans** variables
- M2: Modèle **avec** toutes les variables

- On teste:

$$\begin{cases} H_0 : M1 \rightarrow Logit[P(Y=1)] = \beta_0 \\ H_1 : M2 \rightarrow Logit[P(Y=1)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \end{cases}$$

- Est-ce que M1 est meilleur que M2 (qualités prédictives)?

Test de significativité globale

- La statistique de test est:
  - RV= (-2.ln(vraisemblance au maximum de M1)] - (-2.ln(vraisemblance au maximum deM2)]

Et suit un Khi-deux à **p** degrés de liberté

- Si RV > $\chi^2(p)$ ⟶ **On rejette Ho**, le modèle 2 est meilleur que le 1, les variables explicatives ont simultanément une influence sur la probabilité d'apparition de l'évènement étudié.

Test de significativité pour une variable

- M1: Modèle **sans** la variable testée β_j
- M2: Modèle **avec** la variable testée β_j

- On teste:

$$\begin{cases} H_0: M1 \rightarrow Logit[P(Y=1)] = & \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ H_1: M2 \rightarrow Logit[P(Y=1)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_k x_k \end{cases}$$

$$\text{C'est-à-dire} : \begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

Test de significativité pour une variable

- Il y a 2 manières d'écrire la statistique de test
  - Sous une loi Normale:

$$U = \frac{\widehat{\beta_j}}{s(\widehat{\beta_j})} \sim N(0,1) \qquad \text{sous H0}$$

  - Sous une loi du Khi-deux:

$$X^2 = \left(\frac{\widehat{\beta_j}}{s(\widehat{\beta_j})}\right)^2 = U^2 \sim X^2(1) \qquad \text{sous H0}$$

Test de significativité pour une variable

- Conclusion
  - Sous une loi Normale:

    Si $|U| > N(0,1)$ (=1,96 à 95%)
  - Sous une loi du Khi-deux:

    Si $U > \chi^2(1)$

→ On rejette HO, le modèle 2 est meilleur que le 1, le paramètre β_j **est significatif**, la variable j a une influence sur la probabilité d'apparition de l'évènement, sachant les autres variables du modèle.

Modification d'effet ou interaction

- On considère le modèle M2:

$$Logit[P(Y = 1|X_1, X_2)] = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_1.X_2$$

- Si β3 est significative, alors X2 modifie l'effet de X1. En effet, dans ce cas:
  - Si X2=0 -> l'effet de X1 est β1
  - Si X2=1 -> l'effet de X1 est β1+β3

Modification d'effet ou interaction

- On teste par le test du rapport de vraisemblance:

$$\begin{cases} \boldsymbol{H_0 : \beta_3 = 0} \\ \boldsymbol{H_1 : \beta_3 \neq 0} \end{cases}$$

- Si on rejette HO ⟶ Il y a modification d'effet
  ⟶ On laisse l'interaction dans le modèle.

- Si on accepte HO ⟶ On retire l'interaction.

Confusion

- On considère 2 modèles a et b:

$$logit[P(Y = 1|X_1)] = \beta_0 + \beta_1.X_1$$

$$logit[P(Y = 1|X_1, X_2)] = \beta_0 + \beta_1.X_1 + \beta_2.X_2$$

- Effet brut de X1: $RC_a = \exp(\beta 1)$ de M_a
- Effet de X1 ajusté à X2: $RC_b = \exp(\beta 1)$ de M_b

- Il y a confusion si $RC_a \neq RC_b$

Confusion

- Variation relative:

$$VR = \frac{\left|\widehat{RC_b} - \widehat{RC_a}\right|}{\widehat{RC_b}}$$

- 10%<k<20%

- Si VR>k ⟶ X2 est un facteur de confusion
- Si VR≤k ⟶ on vérifie β2 =0. Si oui, on retire X2 de l'étude.

IV. Adéquation du modèle

Principe

- Déterminer la qualité d'ajustement du modèle aux données.
- Si l'ajustement est correct, les valeurs prédites seront proches des valeurs observées.

Test de Hosmer et Lemeshow

- Regroupement des probabilités prédites $\widehat{y_i}$ par le modèle en dix groupes (déciles).

- Pour chaque groupe, on observe l'écart entre les valeurs prédites et observées. L'importance de la distance entre ces valeurs est évaluée grâce une statistique du Khi-deux à 8 ddl qui teste:

$$\begin{cases} H_0 : Distance\ faible \\ H1 : Distance\ élevée \end{cases}$$

Tableau de contingence

	Malade (yi=1)	Non Malade (yi=0)	Total
Prédit malade ($\widehat{y_i}$ = 1)	**a**	**c**	a+c
Prédit non malade ($\widehat{y_i}$ = 0)	**b**	**d**	b+d
Total	a+b	c+d	n

- Ce tableau permet de connaitre le nombre de bonnes et de mauvaises prédictions par rapport à un seuil « s » (fixé généralement à 50%)

$$Nb_b = \frac{a+d}{n} \qquad Nb_m = \frac{c+b}{n}$$

Tableau de contingence

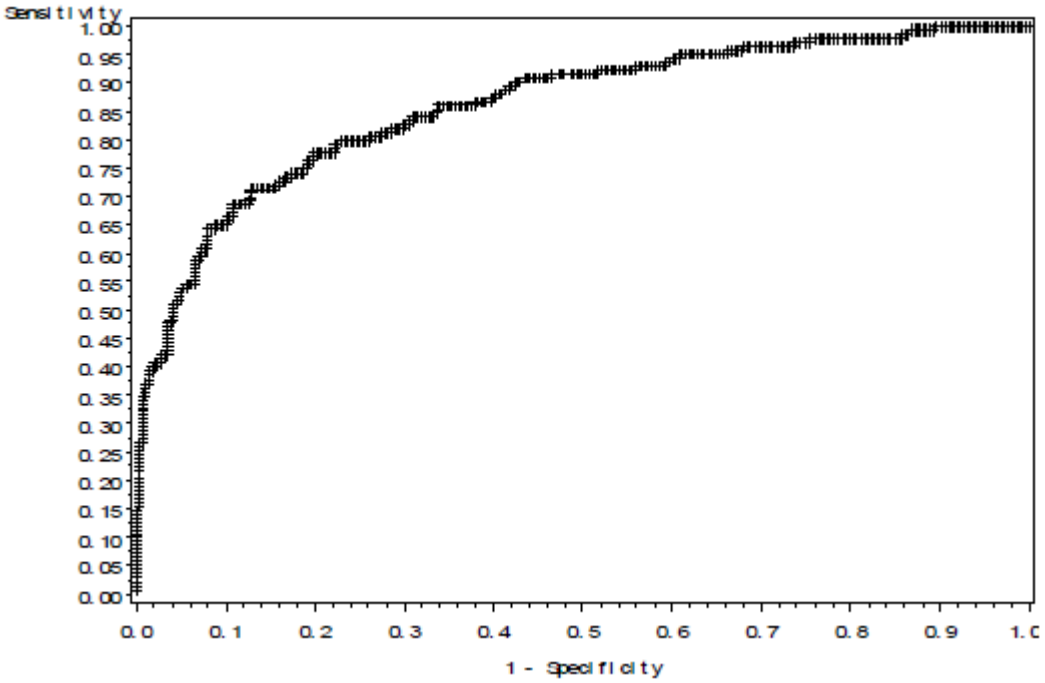
	Malade (yi=1)	Non Malade (yi=0)	Total
Prédit malade ($\widehat{y_i}$ = 1)	**93**	**31**	124
Prédit non malade ($\widehat{y_i}$ = 0)	**50**	**257**	307
Total	143	288	431

- Nbb: 93+257/431=81,2%
- Nbm: 50+31/431= 18,8%

- Sensibilité: Se: 93/143 = 65%
  Sp: 257/288 = 89,2%

Courbe ROC

- S_e en fonction de 1-S_p



- L'aire sous la courbe:

=0,5	Aucune discrimination
]0,5;0,7[	Discrimination faible
[0,7;0,8[	Discrimination acceptable
[0,8;0,9[	Discrimination excellente
[0,9;1]	Discrimination parfaite

V. Application

Description des données

- **REPRISE** : reprise de consommation de drogues avant la fin prévue du programme de traitement (0=non ; 1=oui)
- **SITE** : site du programme (0=A, 1=B)
- **AGE** : âge à l'inclusion
- **BECK** : score de dépression de BECK à l'inclusion (de 0.0 (normal) à 54.0 (dépression)
- **IVHX** : histoire d'utilisation de drogues par voie intraveineuse à l'inclusion (1=jamais ; 2=ancienne ; 3=récente)
- **NBTRAIT** : nombre de traitements anti-drogue précédemment suivis (de 0 à 40)
- **RACE** : race (0=blanche, 1=autre)
- **DUREE** : durée du traitement attribuée par tirage au sort à l'inclusion (0=courte ; 1=longue)

Description des données

Variables		moyenne (écart-type)
Age à l'inclusion		32,38 (6,19)
Score de dépression de Beck à l'inclusion		17,37 (9,33)
Nombre de traitements anti-drogue précédemment suivis		4,54 (5,48)
Variables		**n (%)**
Histoire d'utilisation de drogues par voie intraveineuse à l'inclusion		
	jamais	223 (38,78)
	ancienne	109 (18,96)
	récente	243 (42,26)
Race		
	blanche	430 (74,78)
	autre	145 (25,22)
Durée du traitement		
	courte	289 (50,26)
	longue	286 (49,74)
Site du programme de traitement		
	A	400 (69,57)
	B	175 (30,43)
Reprise de consommation de drogues avant la fin prévue du programme de traitement		
	oui	428 (74,43)
	non	147 (25,57)

Régression logistique multiple

- Hypothèse de « **linéarité du logit** » : il existe une relation linéaire entre le Logit du risque et la variable X.

- Estimation du 1er modèle :

M1 : logit P [REPRISE=1|AGE]=β0+β1*AGE

Analyse des estimations de la vraisemblance maximum

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept	1	1.6602	0.5111	10.5524	0.0012
AGE	1	-0.0182	0.0153	1.4027	0.2363

Régression logistique multiple

- Estimation du 2ème modèle :

M2 : logit P [REPRISE=1|AGE]=β0+β1*AGE(2) + β2*AGE(3) + β4*AGE(4)

Estimations des rapports de cotes

Effet	Point Estimate	95% Limites de confiance de Wald	
age1 2 vs 1	1.370	0.791	2.373
age1 3 vs 1	0.661	0.396	1.104
age1 4 vs 1	1.168	0.687	1.988

- Aucune tendance à la diminution

- Hypothèse de linéarité du logit non vérifiée →Utilisation de la variable AGE en catégorielle

Régression logistique multiple

```
proc logistic data=TP2.donnees  descending;
class IVHX (ref='1') / param=ref;
class age1 (ref='1') / param=ref;
model REPRISE = SITE RACE AGE1 BECK IVHX NBTRAIT DUREE;
run;
```

- Option **descending** : elle inverse l'ordre d'affichage des modalités de la variable dépendante.

- La commande `class IVHX(ref='1') / param=ref;`
demande à SAS de créer des variables indicatrices pour les variables catégorielles IVHX et AGE en prenant comme classe de référence le groupe IVHX=1 et AGE=1.

- **MODEL var_dep = var_indep </ options>;**

IV. Application

Profil de réponse

Valeur ordonnée	REPRISE	Fréquence totale
1	1	428
2	0	147

Probability modeled is REPRISE='1'.

Informations sur le niveau de classe

Classe	Valeur	Variables de création		
IVHX	1	0	0	
	2	1	0	
	3	0	1	
age1	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

État de convergence du modèle

Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	655.729	639.037
SC	660.083	686.935
-2 Log L	653.729	617.037

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	36.6915	10	<.0001
Score	35.1703	10	0.0001
Wald	32.6578	10	0.0003

IV. Application

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
SITE	1	0.8178	0.3658
RACE	1	1.8883	0.1694
age1	3	10.6379	0.0139
BECK	1	0.0248	0.8748
IVHX	2	5.6653	0.0589
NBTRAIT	1	4.8510	0.0276
DUREE	1	5.3105	0.0212

Le Système SAS

The LOGISTIC Procedure

Analyse des estimations de la vraisemblance maximum

Paramètre		DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept		1	1.0966	0.3291	11.1038	0.0009
SITE		1	-0.1991	0.2201	0.8178	0.3658
RACE		1	-0.3104	0.2259	1.8883	0.1694
age1	2	1	0.1798	0.2906	0.3826	0.5362
age1	3	1	-0.6876	0.2768	6.1717	0.0130
age1	4	1	-0.2694	0.3024	0.7940	0.3729
BECK		1	0.00171	0.0108	0.0248	0.8748
IVHX	2	1	0.4825	0.2859	2.8470	0.0915
IVHX	3	1	0.5668	0.2540	4.9785	0.0257
NBTRAIT		1	0.0562	0.0255	4.8510	0.0276
DUREE		1	-0.4628	0.2008	5.3105	0.0212

Estimations des rapports de cotes

Effet		Point Estimate	95% Limites de confiance de Wald	
SITE		0.819	0.532	1.262
RACE		0.733	0.471	1.141
age1	2 vs 1	1.197	0.677	2.116
age1	3 vs 1	0.503	0.292	0.865
age1	4 vs 1	0.764	0.422	1.382
BECK		1.002	0.981	1.023
IVHX	2 vs 1	1.620	0.925	2.837
IVHX	3 vs 1	1.763	1.071	2.900
NBTRAIT		1.058	1.006	1.112
DUREE		0.630	0.425	0.933

Régression logistique multiple

- Sélection des variables : Procédure descendante manuelle

On élimine la variable avec la p-value la plus élevé

1. On enlève la variable BECK (p-value=0.8748 qui est la plus élevée)

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
SITE	1	0.8178	0.3658
RACE	1	1.8883	0.1694
age1	3	10.6379	0.0139
BECK	1	0.0248	0.8748
IVHX	2	5.6653	0.0589
NBTRAIT	1	4.8510	0.0276
DUREE	1	5.3105	0.0212

Régression logistique multiple

2. On ré-estime le modèle sans cette variable et on élimine la variable avec une p-value > 0.05 … etc

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
age1	3	10.4712	0.0150
IVHX	2	8.3230	0.0156
NBTRAIT	1	5.3710	0.0205
DUREE	1	5.5454	0.0185

Régression logistique multiple

Estimations des rapports de cotes

Effet		Point Estimate	95% Limites de confiance de Wald	
age1	2 vs 1	1.152	0.655	2.024
age1	3 vs 1	0.497	0.290	0.852
age1	4 vs 1	0.715	0.399	1.283
IVHX	2 vs 1	1.676	0.963	2.917
IVHX	3 vs 1	1.968	1.223	3.167
NBTRAIT		1.061	1.009	1.115
DUREE		0.625	0.422	0.924

- RC (Age 2 VS 1) = 1.152 → Avoir entre 28 et 33 ans **augmente** la probabilité de reprise de drogue par rapport à un individu ayant un âge inférieur à 28 ans.

- RC (DUREE) = 0.625 → Un individu qui a une durée de traitement longue **diminue** sa probabilité de reprise de drogue, ajusté sur les autres variables explicatives.

Etude de l'interaction entre deux variables

```
proc logistic data=TP2.donnees  descending;
class age1 (ref='1') / param=ref;
model REPRISE = NBTRAIT AGE1 NBTRAIT*AGE1;
run;
```

- La variable AGE modifie-t-elle l'effet de la variable NBTRAIT sur la variable dépendante REPRISE ?

Etude de l'interaction entre deux variables

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
NBTRAIT	1	3.0959	0.0785
age1	3	5.2675	0.1532
NBTRAIT*age1	3	7.9030	0.0481

- On rejette Ho, l'interaction entre AGE et NBTRAIT est significative.
  AGE modifie donc la variable NBTRAIT sur la variable dépendante REPRISE

→ On garde le terme d'interaction. Il y a modification d'effet

Facteur de confusion

- On souhaite déterminer si la durée du traitement (variable DUREE) modifie l'effet du nombre de traitement anti-drogue suivis (variable NBTRAIT) sur le risque de reprise de drogue (variable REPRISE).

1. On vérifie que la variable DUREE ne modifie pas l'effet de NBTRAIT sur la variable dépendante REPRISE.

Analyse des estimations de la vraisemblance maximum

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept	1	0.8442	0.2056	16.8510	<.0001
NBTRAIT	1	0.1195	0.0453	6.9753	0.0083
DUREE	1	-0.1943	0.2660	0.5337	0.4650
NBTRAIT*DUREE	1	-0.0689	0.0533	1.6727	0.1959

Facteur de confusion

2. On considère un 1er modèle M1 :
logit P [REPRISE | NBTRAIT, DUREE] = β0 + β1*NBTRAIT + β2*DUREE

Et un 2ème modèle M2 :
logit P [REPRISE|NBTRAIT]=β0+β1*NBTRAIT

Estimations des rapports de cotes

Effet	Point Estimate	95% Limites de confiance de Wald	
NBTRAIT	1.077	1.026	1.130
DUREE	0.647	0.442	0.948

Estimations des rapports de cotes

Effet	Point Estimate	95% Limites de confiance de Wald	
NBTRAIT	1.078	1.027	1.131

Facteur de confusion

3. On calcule la variation relative $VR = \frac{\left|\hat{RC}_a - \hat{RC}_b\right|}{\hat{RC}_a} =$ (1.077 – 1.078) / (1.077) = 0.0009

La durée du traitement n'est pas un facteur de confusion. Il ne faut pas en tenir compte dans la mesure d'association entre le nombre de traitement anti-drogue suivis et la reprise ou non de drogues.

On retient le modèle M2 :

logit P [REPRISE=1 | NBTRAIT] = $\beta_0 + \beta_1$*NBTRAIT

Adéquation du modèle

```
proc logistic data=TP2.donnees  descending;
/* attention aux valeurs manquantes*/
class IVHX (ref='1') / param=ref ;
class AGE1 (ref='1') / param=ref ;
/*création de 2 variables indicatrices pour la variable IVHX*/
model REPRISE=IVHX NBTRAIT DUREE AGE1 / lackfit
outroc=croc;
run;
```

Adéquation du modèle

Test d'adéquation d'Hosmer et de Lemeshow

Khi 2	DF	Pr > Khi 2
2.1472	8	0.9762

- On accepte Ho → Le modèle est adéquat

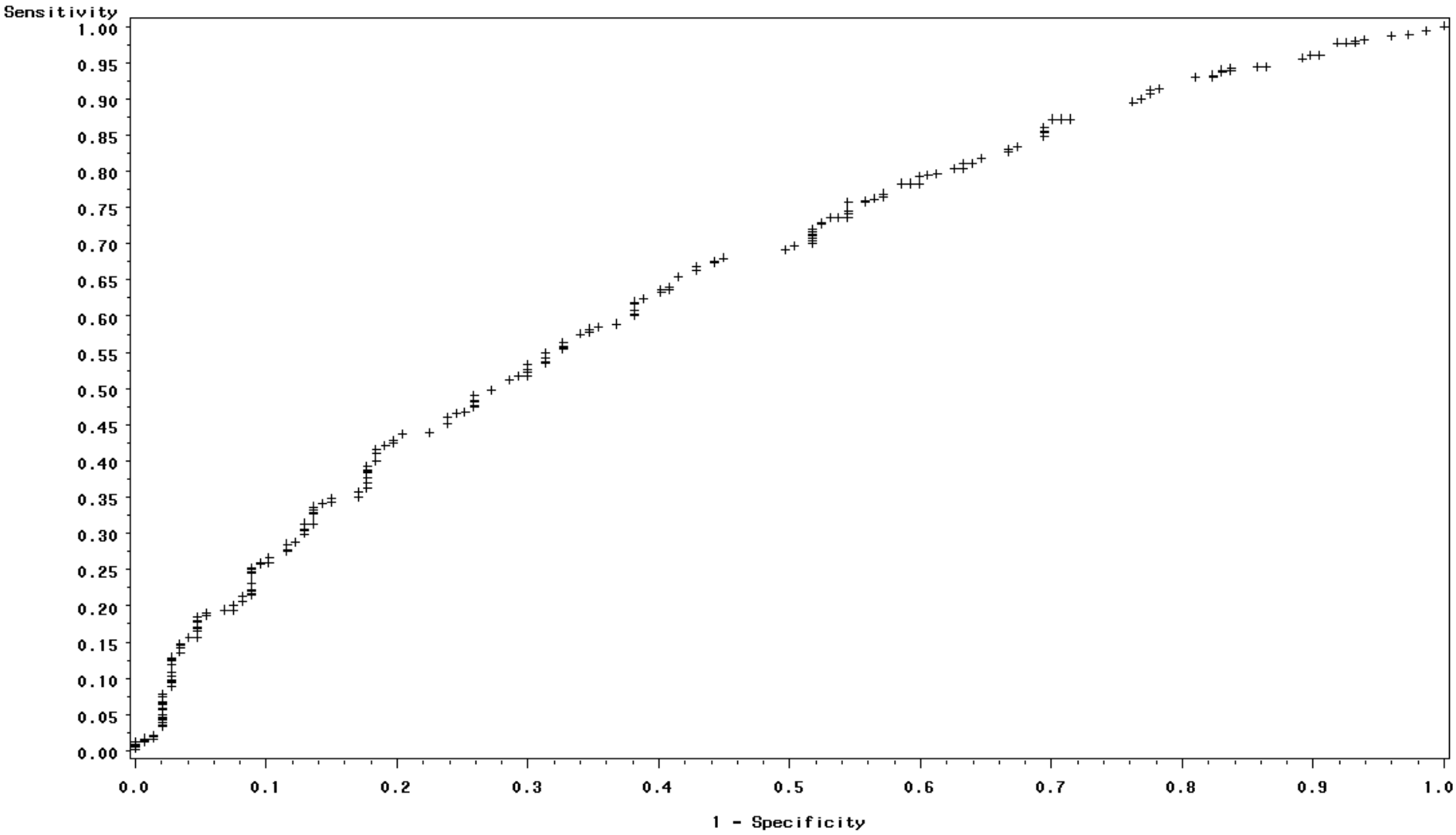
Pouvoir discriminant du modèle

```
/* tracé de la courbe ROC*/
proc gplot data=croc; /*on utilise la table crée
précédemment*/
plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by 0.05;
run;
```

Association des probabilités prédites et des réponses observées

Percent Concordant	65.5	Somers' D	0.320
Percent Discordant	33.5	Gamma	0.323
Percent Tied	1.1	Tau-a	0.122
Pairs	62916	c	0.660

IV. Application



Conclusion

- Variable endogène Y binaire.
- Variable exogène X quantitative ou qualitative
  - Si quantitative, vérifier l'hypothèse de linéarité.
- Les paramètres ne sont **pas interprétables**
  - Il faut calculer les RC=exp(βk) et les comparer à 1
- Les tests sont tous basés sur la test **du rapport de vraisemblance.**
- **Adéquation** du modèle: On mesure l'écart entre les valeurs prédites et observées.