Régression logistique

Application à la base de données du TITANIC

Bérangère BERTHO & Laura SEBILLE

Sommaire

- Partie 1 : Généralités
- Partie 2 : Interprétations
- Partie 3 : Pour un meilleur modèle…

Partie 1 : Généralités

- 1. Principe et estimation
- 2. Bonne ou mauvaise régression ?
- 3. Significativité des coefficients

Introduction

Introduction

- Objectif : Prédire les valeurs prises par Y définie dans $\{y_1, y_2, \ldots, y_K\}$
- Y prend deux modalités : {+,−} ou {1,0}
- Echantillon Ω de taille n
- La valeur prise par Y pour un individu w est notée Y(w)

Introduction

- J descripteurs $\{X_1, X_2, \ldots, X_J\}$
- Le vecteur de valeurs pour un individu w s'écrit $(X_1(w), X_2(w), \ldots, X_J(w))$
- $P[Y(w) = 1] = p(w)$
- Probabilité a posteriori :
  $P[Y(w) = 1|X(w)] = \pi(w)$
- LOGIT d'un individu w

Introduction

- Données
- 2201 observations
- 3 variables prédictives
- Objectif : prédire la survie (ou le décès) d'un passager du Titanic
- Y = 1 si survie, 0 sinon
- Classe : 0 à 3
- Age : 0 (enfant) ou 1 (adulte)
- Sexe : 0 (femme) ou 1 (homme)

Principe et estimation

Principe et estimation : Un cadre bayesien pour l'apprentissage supervisé

- $Y = f(X, \alpha)$

  α est le vecteur des paramètres de la fonction

- Comment évaluer la qualité de la modélisation ?
  - Mesurer la qualité de prédiction dans la population Ω^{pop}
  - Erreur théorique

Principe et estimation : Un cadre bayesien pour l'apprentissage supervisé

- $P[Y(w) = y_k \mid X(w)]$
- $Y_k^* = \arg\max_k P[Y(w) = y_k \mid X(w)]$

- Application : Titanic
  - Survie = f(Classe)

	CLASSE				
SURV	0	1	2	3	Total
0	76%	38%	59%	75%	68%
1	24%	62%	41%	25%	32%
Total	100%	100%	100%	100%	100%

29 % de chance de se tromper

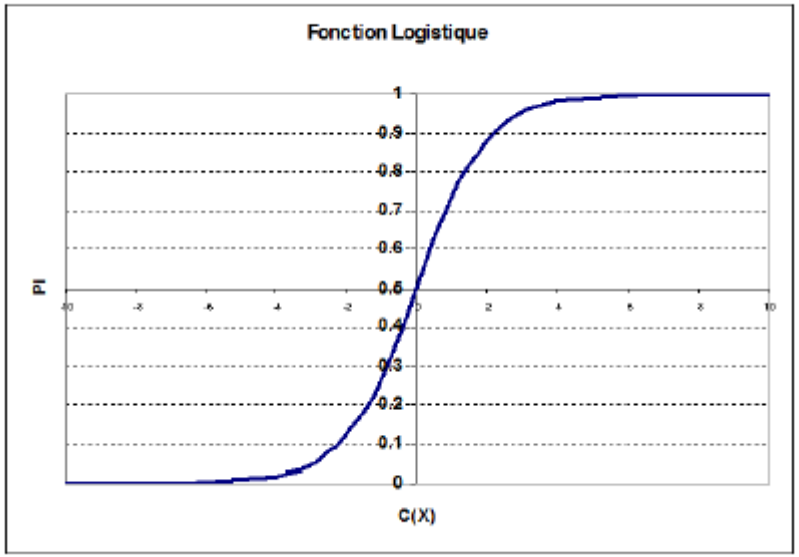
- Problèmes :
  - Lourdeur des calculs si nombreuses variables
  - Impossible à exploiter si faibles effectifs
  - Impossible à utiliser tels quels les descripteurs continus

Principe et estimation : Hypothèse fondamentale de la régression logistique

- La régression logistique produit un séparateur linéaire
- $P(Y = y_k \mid X)$
- Régression logistique est une méthode semi-paramétrique

Principe et estimation : Le modèle LOGIT

- Fonction logistique



- $P(Y = 1 \mid X) + P(Y = 0 \mid X) = 1$

Principe et estimation : Estimation par maximum de vraisemblance

- Vraisemblance
- Log-vraisemblance
- Déviance
- Optimisation

- Application : Titanic
- Occurrence de la survie :

$$\text{logit} = 2{,}05 + 0{,}2*1_{\text{Classe}=0} + 1{,}06*1_{\text{Classe}=1} + 0{,}04*1_{\text{Classe}=2} - 0{,}72*1_{\text{Classe}=3} - 1{,}06*1_{\text{age}=\text{adulte}} - 2{,}42*1_{\text{sexe}=\text{Homme}}$$

Principe et estimation : Première évaluation de la régression

- Modèle trivial M_0
- Vérifier que le modèle fait mieux que le modèle trivial
- R^2 de Mac-Fadden :

  $R^2_{MF} = 1 - LL_M / LL_0 = 0{,}20$

Bonne ou mauvaise régression ?

Bonne ou mauvaise régression ? : La matrice de confusion

- a sont les **vrais** positifs : individus qui ont été classés positifs et qui le sont réellement
- c sont les **faux** positifs : classés positifs alors qu'ils sont négatifs

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Bonne ou mauvaise régression ? : La matrice de confusion

- Taux d'erreur : nombre de mauvais classement rapporté à l'effectif total
- Taux de succès : probabilité de bon classement du modèle
- Sensibilité : capacité du modèle à retrouver les positifs
- Précision : proportion de vrais positifs parmi les individus classés positifs
- Spécificité : proportion de négatifs détectés

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Bonne ou mauvaise régression ? : La matrice de confusion

- Le taux de faux positifs (TFP) : proportion de négatifs qui ont été classés positifs

- La F-Mesure: moyenne harmonique entre la sensibilité et la précision

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	Total
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Bonne ou mauvaise régression ? : La matrice de confusion

- Survie = f(age, sexe, classe)

	Prédiction		
SURV	Décès	Survie	Total général
Décès	1364	126	1490
Survie	362	349	711
Total général	1726	475	2201

Indicateurs	
Vrais positifs	349
Faux positifs	126
Taux d'erreur	22%
Taux de succès	78%
Sensibilité	92%
Précision	79%
Spécificité	49%
F-Mesure	0,85
Rapp. Vraisemblance	1,80

- R^2_e = **0,31**

  Ce modèle est meilleur que le modèle trivial

Bonne ou mauvaise régression ? : La matrice de confusion

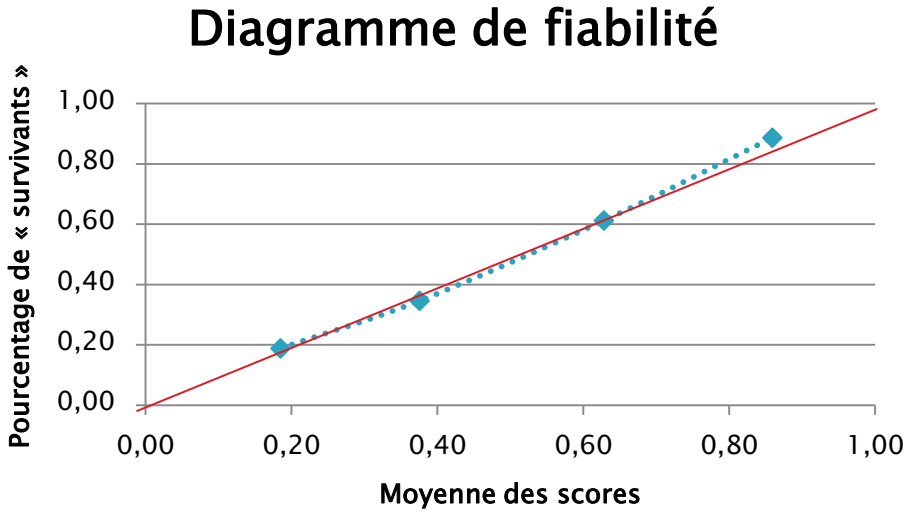
- Attention lors de son utilisation :
  - Se repose sur les prédictions sans tenir compte des probabilités

Probabilité estimée 0.495	Probabilité estimée 0.505
0	**1**

  - Le classement dans le groupe le plus important est toujours favorisé

Bonne ou mauvaise régression ? : Diagramme de fiabilité

- Confronter les probabilités estimées et celles observées



- Ici, les points sont proches de la droite, ainsi les scores sont bien calibrés

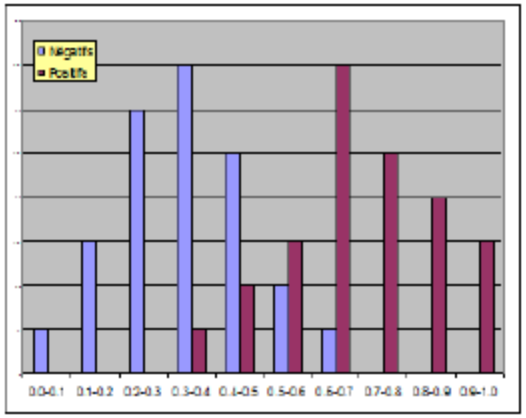
Bonne ou mauvaise régression ? : Test de Hosmer-Lemeshow

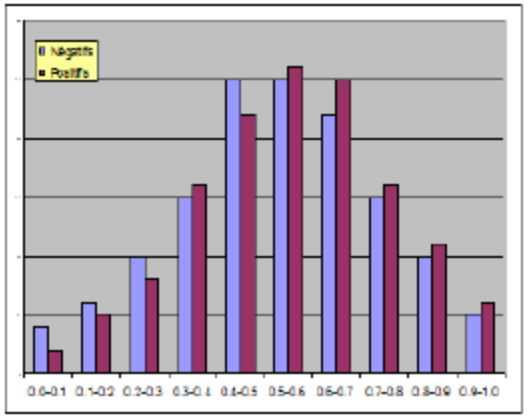
- Repose sur la même logique que le diagramme de fiabilité
- H0 : Le modèle n'apporte rien
  H1 : Le modèle logistique apporte de l'information
- Statistique de test $C \sim Khi^2_{(G-2)}$

- $1592 > Khi^2_{(0,95\ ;\ G-2)} = 0,35$
- On rejette H0 : Le modèle est validé

Bonne ou mauvaise régression ? : Le test de Mann-Whitney

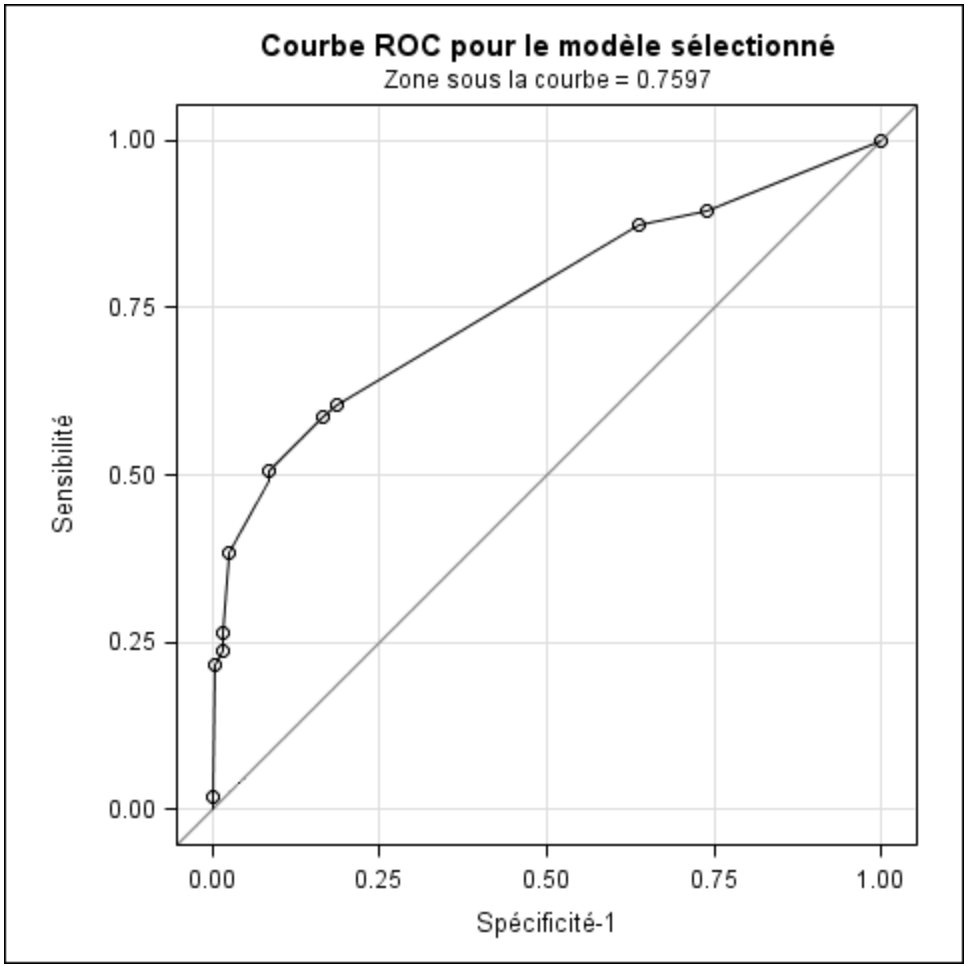
- H0 : Les scores des survivants ne sont pas significativement différents des scores des décédés





- Statistique de test Z ~ N(0;1)
- |-13,43| > U(95)=1,96

Bonne ou mauvaise régression ? : la courbe ROC



Test de significativité des coefficients

Test de significativité des coefficients : Quoi et comment tester ?

- Différentes hypothèses à tester pour valider le modèle :
  - $H_0 : a_j = 0$
  - $H_0 : a_j = a_{j+1} = \ldots = a_{j+q} = 0$
  - $H_0 : a_1 = a_2 = \ldots = a_J = 0$

- 2 approches pour les tests :
  - Rapport de vraisemblance
  - Test de Wald

Test de significativité des coefficients : Tests fondés sur le rapport de vraisemblance

- 2 modèles
  - $M_r : Y = a_0+a_1X_1 +\ldots+a_sX_s+a_{s+1}X_{s+1}+\ldots+a_rX_r$
  - $M_s : Y = a_0+a_1X_1 +\ldots+a_sX_s$
- Comparaison des vraisemblances des modèles emboités
- $H_0 : a_{s+1}=\ldots=a_r = 0$
- Statistique de test LR $\sim Khi^2_{(r-s)}$

Test de significativité des coefficients : Tests de Wald

- Repose sur des propriétés asymptotiques de l'estimateur

- $H_0 : a_1 = \ldots = a_j = 0$

- Statistique de test $W_j \sim Khi^2_{(j)}$

➡ On privilégie le test du rapport de vraisemblance pour les petites bases de données car il est plus puissant.

➡ Pour les grosses bases de données, on privilégie le test de Wald qui est moins gourmand en ressources.

Partie 2 : Interprétations

- 1. Interprétation des coefficients
- 2. Les interactions

Interprétation des coefficients

Interprétation des coefficients : Risque relatif, odds, odds-ratio

- Risque relatif : surcroit de chance d'être positif du groupe exposé par rapport au groupe témoin
- Odds : rapport de probabilités dans un groupe
- Odds-ratio : rapport entre l'odds du groupe exposé et l'odds du groupe témoin

Interprétation des coefficients : Le cas de la régression simple

- Survie = f(age)
- Odds = 0,45
  - Les adultes avaient 2,2 fois plus de risque de décéder que de rester en vie.
- Odds-ratio = 0,41
  - On a 2,4 fois plus de chance de survivre en étant enfant qu'en étant adulte.
- Intervalle de confiance de l'OR : [0,28 ; 0.61]
  - Lien significatif entre l'âge et la survie

Interprétation des coefficients : Le cas de la régression simple

- Survie = f(classe)

Estimation par l'analyse du maximum de vraisemblance

Paramètre		DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-1.0873	0.0867	157.3831	<.0001
CLASS	0	1	-0.0678	0.1171	0.3356	0.5624
CLASS	1	1	1.5965	0.1436	123.5199	<.0001
CLASS	2	1	0.7400	0.1482	24.9200	<.0001

Estimation des odds-ratio

Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
CLASS 0 vs 3	0.934	0.743	1.175
CLASS 1 vs 3	4.936	3.725	6.541
CLASS 2 vs 3	2.096	1.567	2.803

Interprétation des coefficients : Le cas de la régression multiple

- Survie = f(sexe, age)

Estimation par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	1.5140	0.2355	41.3294	<.0001
AGE	1	-0.5564	0.2276	5.9770	0.0145
SEX	1	-2.2940	0.1199	365.7942	<.0001

Estimation des odds-ratio

Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
AGE	0.573	0.367	0.896
SEX	0.101	0.080	0.128

Les interactions

- Exemple :
  - X_1 = adulte et X_2 = sexe masculin
  - $Z = X_1 * X_2$ = 1 si adulte ET sexe masculin
  - Z = 0
    - Si adulte et de sexe féminin
    - Si enfant et de sexe masculin
    - Si enfant et de sexe féminin

- Application aux données du Titanic

Estimation par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	0.4990	0.3075	2.6338	0.1046
AGE	1	0.5654	0.3269	2.9911	0.0837
SEX	1	-0.6870	0.3970	2.9953	0.0835
AGE*SEX	1	-1.7465	0.4167	17.5661	<.0001

Les interactions : Calcul de l'odds-ratio

- Logit = 0,5 + 0,57 x Adulte – 0,69 x Homme – 1,75 x Homme * Adulte
- $\Delta_{logit}(sexe) = -0,69 - 1,75 \times 1 = -2,44$
- $e^{-2,44} = 0.09$
  - Un homme adulte a 11 fois plus de risque de se noyer qu'une femme adulte
- Un garçon a 3 fois plus de chance de survivre qu'un homme

Partie 3 : Pour un meilleur modèle

- 1. Sélection de variables
- 2. Analyse des résidus

La sélection de variables

La sélection de variables : Pourquoi la sélection ?

- Moins il y aura de variables, plus facile sera l'interprétation
- Le déploiement sera facilité
- Plus de chance que le modèle soit robuste
  - Principe du Rasoir d'Occam

- La sélection manuelle est à préférer (à condition d'être expert dans le domaine)

La sélection de variables : Pourquoi la sélection ?

- Procédures pas-à-pas :
  - Sélection FORWARD
  - Sélection BACKWARD
  - Méthode STEPWISE

La sélection de variables : Sélection par optimisation

- Objectif : minimiser un des 2 critères :
  - AIC = −2LL + 2 x (J+1)
  - BIC = −2LL + ln(n) x (J+1)
    - Avec −2LL la déviance
    - J+1 le nombre de paramètres à estimer
    - J le nombre de variables explicatives

- Survie = f(classe, sexe, age)

Estimation par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	2.1448	0.2766	60.1435	<.0001
class1	1	0.9022	0.1498	36.2820	<.0001
class3	1	-0.8634	0.1352	40.7638	<.0001
AGE	1	-1.0315	0.2412	18.2893	<.0001
SEX	1	-2.3813	0.1338	316.9023	<.0001

Estimation des odds-ratio

Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
class1	2.465	1.838	3.306
class3	0.422	0.324	0.550
AGE	0.356	0.222	0.572
SEX	0.092	0.071	0.120

- Survie = f(classe, sexe, age)
  - Avec interaction

Estimation par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	3.1208	0.3278	90.6106	<.0001
AGE	1	-0.9709	0.2278	18.1663	<.0001
SEX	1	-3.5167	0.2453	205.5231	<.0001
class1	1	0.7760	0.1632	22.6144	<.0001
class3	1	-2.4654	0.2834	75.6842	<.0001
SEX*class3	1	2.1465	0.3088	48.3280	<.0001

Proportion de survivants	CLASSE			
SEXE	Equipage	1ère classe	2e classe	3e classe
Femme	87%	97%	88%	46%
Homme	22%	34%	14%	17%

Analyse des résidus

Analyse des résidus

- Deux tests principaux :
  - Résidus de déviance (Ecart)
  - Résidus de Pearson

H0 : Bon ajustement du modèle aux données
H1 : Qualité du modèle insuffisante