Régression logistique

Michaël Genin

Université de Lille 2

EA 2694 - Santé Publique : Epidémiologie et Qualité des soins

michael.genin@univ-lille2.fr

Université de Lille 2 DROIT ET SANTÉ

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Plan

1. Contexte général
2. Introduction à la régression logistique
   - Contexte
   - Exemple introductif
   - Modèle logistique
3. Estimation des coefficients et tests
   - Estimation des coefficients
   - Tests dans le modèle
4. Interprétation des coefficients
   - Rappel sur l'odds-ratio
   - Lien entre OR, modèle Logit et coefficient de la régression
   - Variable explicative quantitative
   - Variable explicative qualitative (+ de 2 modalités)
5. Validation du modèle
   - Pouvoir discriminant du modèle
   - Calibration du modèle
6. Exemple

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Contexte : Deux familles de méthodes de classification

Classification non-supervisée (clustering)

- Partitionner les observations en groupes différents (classes, catégories) mais les plus homogènes possible au regard de variables décrivant les observations.
- **Le nombre de classes n'est pas connu à l'avance**
- Méthodes : Classification hiérarchique, K-plus-proches voisins, Classification bayésienne naïve

Classification supervisée (discrimination)

- Obtenir un critère de séparation afin de prédire l'appartenance à une classe ($Y = f(X) + \epsilon$).
- **Le nombre de classes est connu à l'avance (Variable à expliquer)**
- Méthodes : **Régression logistique**, Analyse discriminante, Arbres de décision, Réseaux de neurones...

Classification supervisée

2 objectifs principaux :

- Etude du lien entre Y (Variable à expliquer : classes) et les X_j (Variables explicatives) $\Rightarrow$ Facteurs prédictifs
- Prédiction (système d'aide à la décision (scores cliniques, crédit scoring, ...)

Différentes procédures :

- 2 classes $\Rightarrow$ Régression logistique
- $>$ 2 classes : Analyse discriminante, Arbres de décision, Réseaux de neurones,...

Point étudié

1 Contexte général

2 Introduction à la régression logistique
- Contexte
- Exemple introductif
- Modèle logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Point étudié

1 Contexte général

2 Introduction à la régression logistique
- Contexte
- Exemple introductif
- Modèle logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Introduction à la régression logistique

Régression logistique = méthode de régression

- Etudier le lien entre une Variable A Expliquer (VAE) **qualitative** Y
- ET
- $\{X_j\}_{j:1\ldots p}$ variables explicatives **quantitatives ou binaires**

Très utilisée en épidémiologie

- Etape liaison (facteurs prédictifs) et ajustement
- Prédiction : Création de scores pronostiques

Introduction à la régression logistique

3 types de régression logistique

- binaire ⇒ VAE binaire (ex : vivant / décés)
- ordinale ⇒ VAE ordinale (ex : stades de cancer)
- multinomiale ⇒ VAE qualitative (ex : types de cancer)

Cours basé uniquement sur la régression logistique binaire car :

- *Reg. Ordinale* : hypothèses complémentaires fortes (proportionnalité entre les modalités de Y)
- *Reg. Multinomiale* : peut être vue comme plusieurs régressions logistiques binaires. L'interprétation des coefficients est plus difficile.

Introduction à la régression logistique

Rappel

En régression linéaire multiple, le modèle est linéaire :

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

Question

Qu'en est-il de la régression logistique ?

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon = \ ?$$

Point étudié

1 Contexte général

2 Introduction à la régression logistique
- Contexte
- Exemple introductif
- Modèle logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle

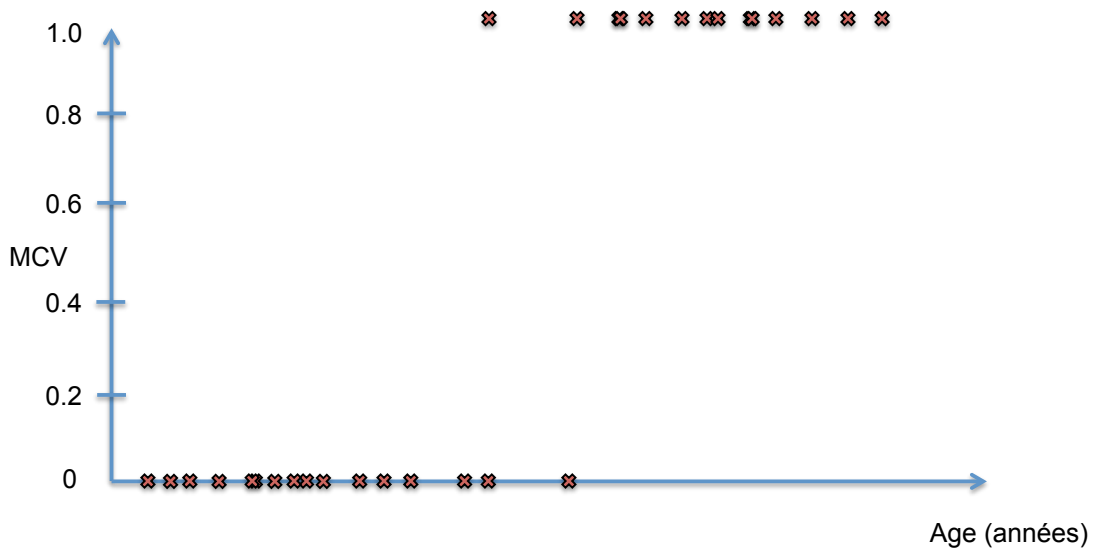
6 Exemple

Introduction à la régression logistique

Exemple introductif

- Y VAE binaire (1 ou 0) $\Rightarrow$ Présence (ou absence) de maladie cardiovasculaire
- Une seule variable explicative quantitative X : l'âge

Représentation graphique

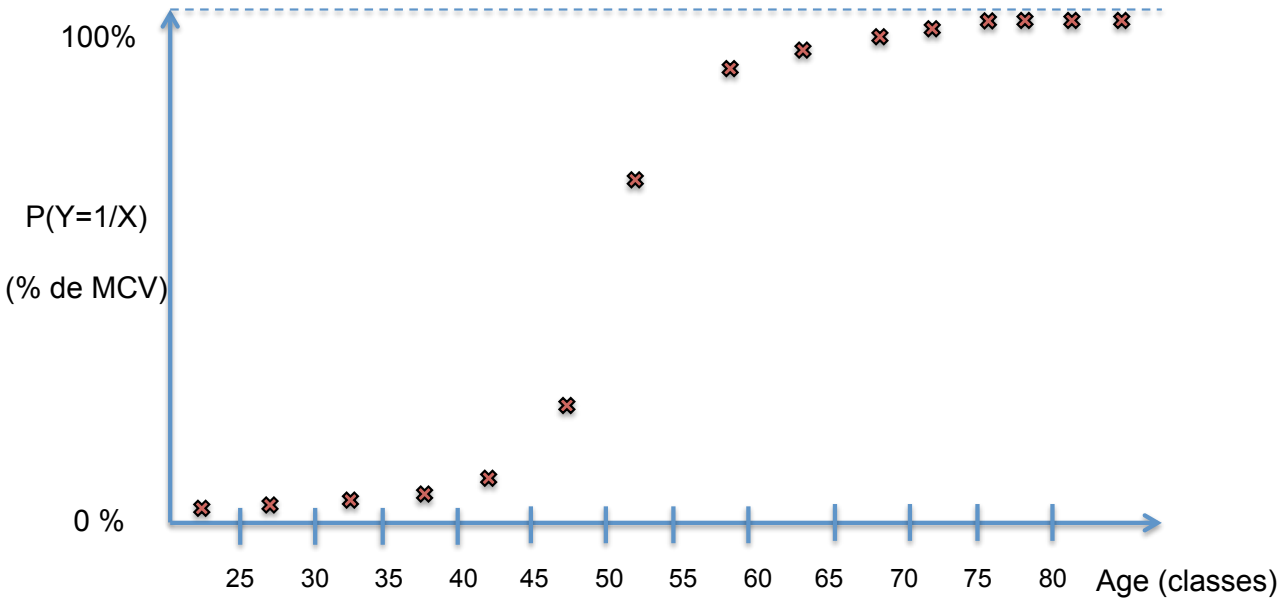


Remarque : Pas vraiment intéressant, pas d'échelle naturelle $\Rightarrow$ VAE qualitative

Introduction à la régression logistique

Idée

Modéliser les modalités de Y (présence ou absence) en termes de % par rapport à X

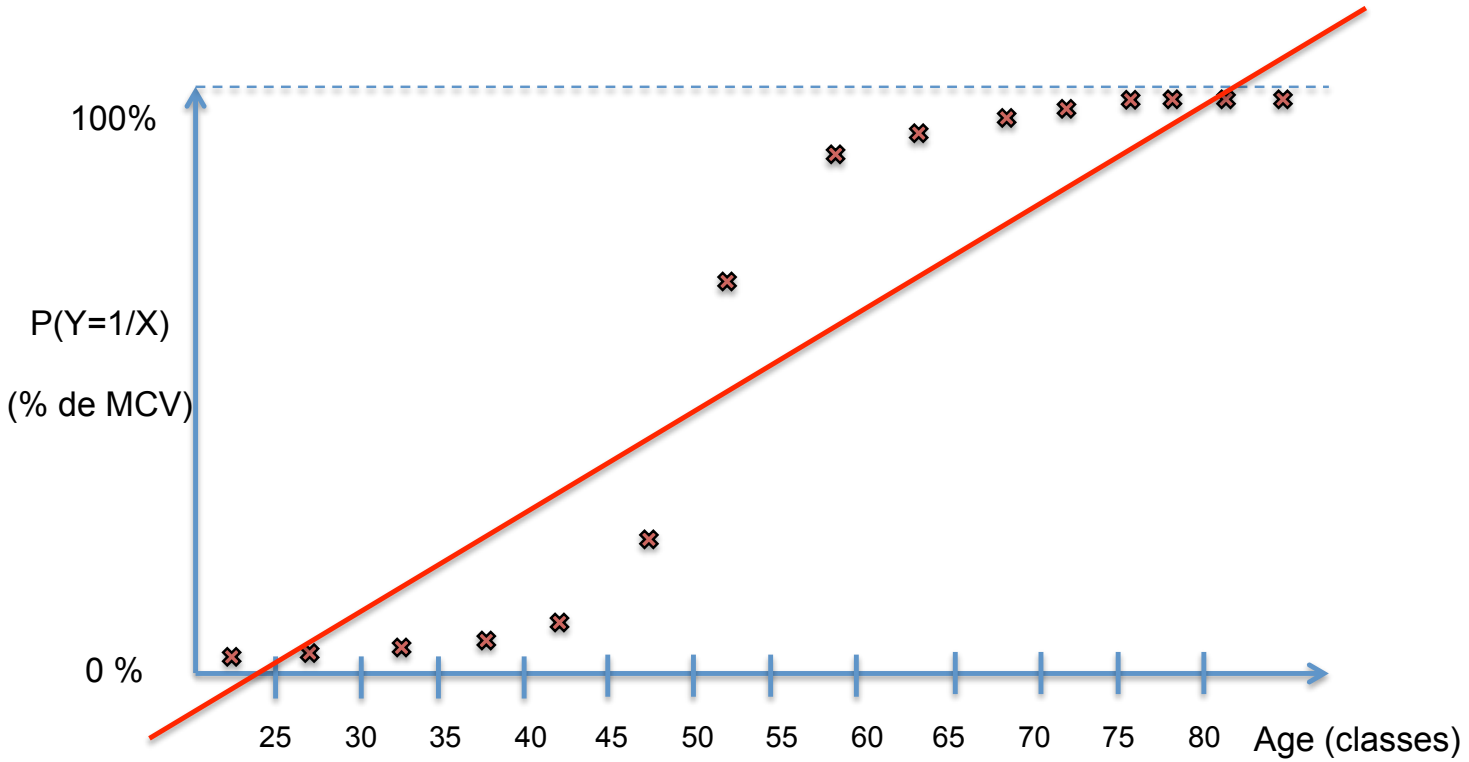


Introduction à la régression logistique

Question

- $\mathbb{P}(Y = 1/X)$ est un attribut numérique

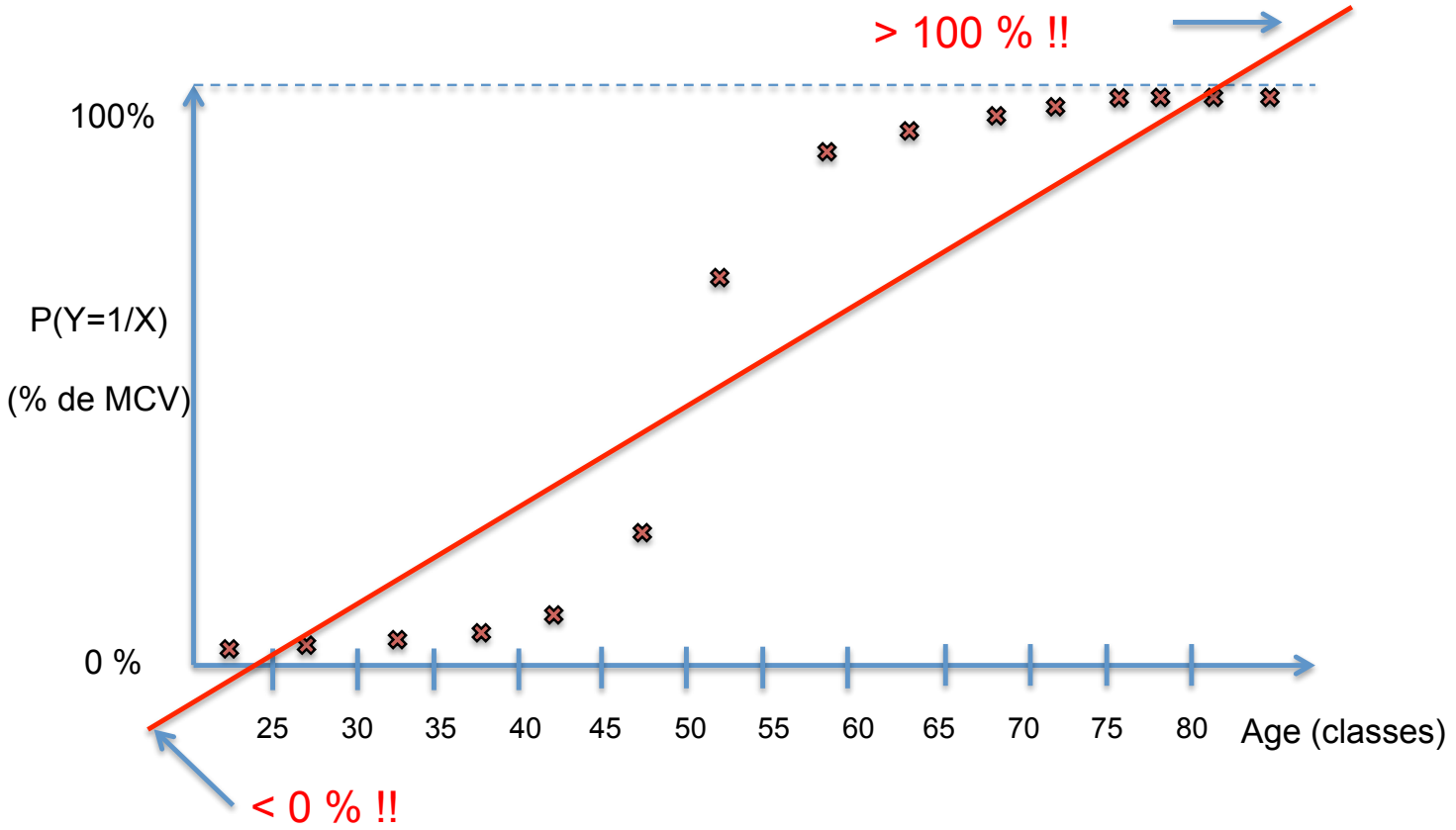
$\Rightarrow$ utilisation d'un modèle linéaire ?



Introduction à la régression logistique

Problème évident

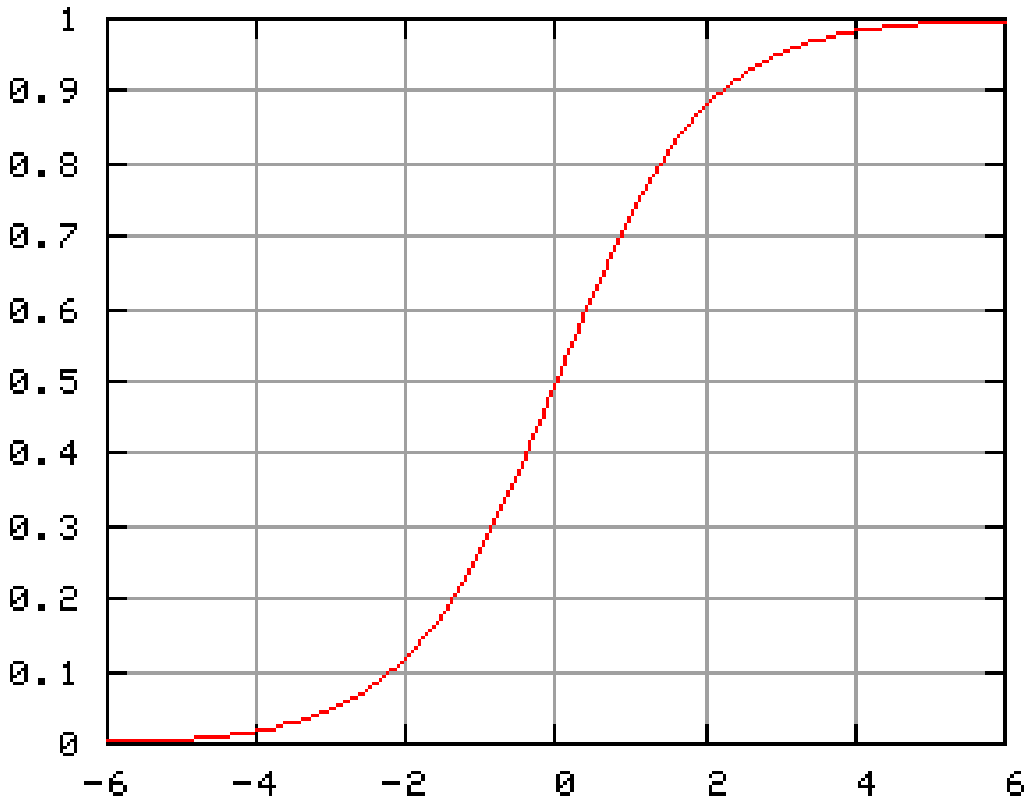
- $\mathbb{P}(Y/X = 1) \in [0, 1]$
- Or si on modélise par une régression linéaire, $\mathbb{P}(Y/X = 1) \in]-\infty; +\infty[$



Introduction à la régression logistique

Nécessité de trouver une nouvelle modélisation (lien non-linéaire)

$\Rightarrow$ Courbe logistique



Point étudié

1 Contexte général

2 Introduction à la régression logistique
- Contexte
- Exemple introductif
- Modèle logistique

3 Estimation des coefficients et tests

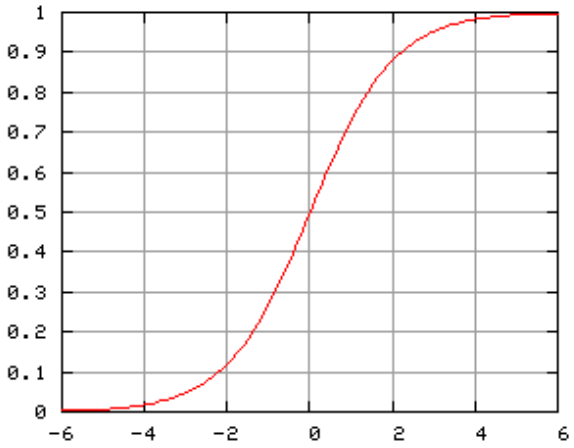
4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Modèle logistique - Courbe logistique / Fonction sigmoïde

Courbe logistique



Fonction sigmoïde :

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

Propriétés intéressantes

$$X \to +\infty \text{ alors } \pi(X) \to 1$$

$$X \to -\infty \text{ alors } \pi(X) \to 0$$

$$\pi(X) \in [0, 1]$$

Modèle logistique - formulation

Modèle logistique à une variable explicative X

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

Remarque sur les erreurs

- $\epsilon = 1 - \pi(X)$ si $Y = 1$
- $\epsilon = -\pi(X)$ si $Y = 0$

Modèle logistique multivarié

$$Y = \mathbb{P}(Y = 1/\{X_j\}) + \epsilon = \pi(\{X_j\}) + \epsilon = \frac{e^{\beta_0+\sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0+\sum_{j=1}^{p} \beta_j X_j}} + \epsilon$$

Modèle logistique - Transformation LOGIT

Transformation LOGIT

$$\text{Logit}[\pi(X)] = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X$$

Intérêts

- Permet de revenir à un modèle linéaire classique
- Interprétation des coefficients du modèle comme une mesure d'association de X par rapport à Y (Notion d'odds-ratio)

Modèle logistique - Construction

Construction du modèle comme en régression linéaire multiple :

- Estimation des coefficients
- Tests
- Interprétation des coefficients
- Mesure d'adéquation et validité du modèle

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests
- Estimation des coefficients
- Tests dans le modèle

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests
- Estimation des coefficients
- Tests dans le modèle

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Estimation des coefficients - Généralités

En régression linéaire multiple $\Rightarrow$ Méthode des moindres carrés ordinaires (MC0)

RLM : Minimisation du critère des MCO

$$\min \sum_{i=1}^{n} (e_i)^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_j))$$

En régression logistique, la MMC ne permet pas d'obtenir une estimation des coefficients. On utilise la

Méthode du maximum de vraisemblance (Maximum likelihood)

Méthode classique qui permet d'estimer les paramètres d'une loi, d'un modèle.

Maximum de vraisemblance (1/5)

Exemple simple : Y binaire (0/1) et une seule variable explicative X quantitative

$$\text{Population} \begin{cases} Y(0/1) \\ X \end{cases} \underset{n-\text{echantillon}}{\longrightarrow} (y_i, x_i)_{i:1\ldots n}$$

Avec pour une observation i :

$$Y_i = \begin{cases} 1 \text{ avec une proba : } \pi(x_i) = \mathbb{P}(Y = 1/X = x_i) \\ 0 \text{ avec une proba : } 1 - \pi(x_i) \end{cases}$$

$$Y_i \sim \mathcal{B}(1, \pi(x_i))$$

avec

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

Maximum de vraisemblance (2/5)

Sur cet échantillon on peut calculer une vraisemblance (probabilité d'observer l'échantillon)

$$L = \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

$$L = \mathbb{P}\left(\{Y_1 = y_1\} \cap \{Y_2 = y_2\} \cap \ldots \cap \{Y_n = y_n\}\right) = \bigcap_{i=1}^{n} \mathbb{P}(Y_i = y_i)$$

Comme les observations sont considérées indépendantes entre elles :

$$L = \bigcap_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i)$$

Maximum de vraisemblance (3/5)

$\mathbb{P}(Y_i = y_i)$ est exprimée de la sorte (Loi de Bernoulli) :

$$\mathbb{P}(Y_i = y_i) = \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

Donc

$$L = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

Or selon le modèle logistique :

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

Aussi

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

Maximum de vraisemblance (4/5)

Méthode du maximum de vraisemblance

Objectifs : trouver $\hat{\beta}_0$ et $\hat{\beta}_1$ qui maximisent la probabilité d'observer l'échantillon (*i.e.* maximisation de la vraisemblance)

$$\max_{\beta_0,\beta_1}(L) = \max_{\beta_0,\beta_1}\left(\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)\right) = \max_{\beta_0,\beta_1}\left(\prod_{i=1}^{n}\mathbb{P}(Y_i = y_i)\right)$$

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

Pour des raisons de simplicité de calcul, on passe généralement par le log

$$\max_{\beta_0,\beta_1}(L) = \max_{\beta_0,\beta_1}\log(L)$$

▶ Détails

Maximum de vraisemblance - (5/5)

Méthode de Newton-Raphson (analyse numérique)

Pour trouver $\hat{\beta}_0$ et $\hat{\beta}_1$ qui maximisent L (ou $\log L$), on a recours aux dérivées partielles :

$$\frac{\partial L}{\partial \beta_0} = 0 \text{ et } \frac{\partial L}{\partial \beta_1} = 0$$

Intérêts

Une fois $\hat{\beta}_0$ et $\hat{\beta}_1$ estimés, on peut calculer, pour tout i, $\hat{\pi}(x_i)$:

$$\hat{\pi}(x_i) = \mathbb{P}(Y_i = 1/X = x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}$$

Ou avec le modèle Logit

$$\text{logit}(\hat{\pi}(x_i)) = \ln\left(\frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests
- Estimation des coefficients
- Tests dans le modèle

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Test global de significativité

Test de Rapport de Vraisemblance (TRV)

- $\mathcal{H}_0$: Pas de liaison entre Y et les $X_j \Leftrightarrow \beta_1 = \beta_2 = \ldots = \beta_p = 0$
- $\mathcal{H}_1$: Le modèle a du sens $\Leftrightarrow$ Au moins 1 $\beta_j \neq 0$

Exemple simple : considérons le cas avec une seule variable explicative X

Principe du TRV

Comparer la vraisemblance L_X (avec variable explicative $(\mathcal{H}_1)$) avec la vraisemblance L_0 sans variable explicative $(\mathcal{H}_0)$.

Intuitivement

Si $L_X > L_0$ alors la variable X apporte à l'estimation de $\mathbb{P}(Y)$

Test global de significativité

Construction de la statistique de test

- L_x vraisemblance avec $X \rightarrow$ déjà calculée
- L_0 vraisemblance sans X (sous $\mathcal{H}_0$)

Sur l'échantillon de taille N on observe : $\begin{cases} card\{y=1\} = n_1 \\ card\{y=0\} = N - n_1 \end{cases}$

Avec $\hat{\pi} = \mathbb{P}(Y=1) = \dfrac{n_1}{N}$ et $1-\hat{\pi} = \mathbb{P}(Y=0) = 1 - \dfrac{n_1}{N}$

Aussi

$$L_0 = \prod_{i=1}^{N} \hat{\pi}^{y_i}[1-\hat{\pi}]^{1-y_i} = \prod_{i=1}^{N} \left(\frac{n_1}{N}\right)^{y_i} \left(1-\frac{n_1}{N}\right)^{1-y_i} = \left(\frac{n_1}{N}\right)^{\sum_{i=1}^{N} y_i} \left(1-\frac{n_1}{N}\right)^{\sum_{i=1}^{N}(1-y_i)}$$

$$L_0 = \left(\frac{n_1}{N}\right)^{n_1} \left(1-\frac{n_1}{N}\right)^{N-n_1}$$

Test global de significativité

Statistique de test du TRV

On montre que sous $\mathcal{H}_0$:

$$D = -2 \ln \left(\frac{L_0}{L_X} \right) \sim \chi^2_{1 \ d.l.l.}$$

Interprétation

- Non rejet de $\mathcal{H}_0$: le modèle n'a pas de sens, X n'apporte rien à l'estimation de $\mathbb{P}(Y)$
- Rejet de $\mathcal{H}_0$: le modèle a du sens, X apporte à l'estimation de $\mathbb{P}(Y)$

Test global de significativité

Extension du test à p variables explicatives X_j

Hypothèses du test

- $\mathcal{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (absence de liaison)
- $\mathcal{H}_1 : \exists$ au moins un $\beta_j \neq 0$ (présence de liaison)

Statistique de test sous $\mathcal{H}_0$

$$D = -2 \ln \left(\frac{L_0}{L_{X_j}} \right) \sim \chi^2_{p \text{ d.l.l.}}$$

Test global / Tests individuels

Dans le cas d'un régression logistique multiple

Si on ne rejette pas $\mathcal{H}_0$ associée au TRV alors STOP

Si on rejette $\mathcal{H}_0$ alors test individuel de chaque coefficient :

Test sur un coefficient - Hypothèses

- $\mathcal{H}_\mathbf{0}$: $\beta_j = 0$ (la variable n'est pas significative dans le modèle)
- $\mathcal{H}_\mathbf{1}$: $\beta_j \neq 0$ (la variable est significative dans le modèle)

Test sur un coefficient - Statistique de test (Test de Wald)

On peut montrer que si $\mathcal{H}_0$ est vraie alors :

$$K = \frac{\hat{\beta}_j^2}{s^2_{\hat{\beta}_j}} \sim \chi^2_1 \text{ ddl}$$

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients
- Rappel sur l'odds-ratio
- Lien entre OR, modèle Logit et coefficient de la régression
- Variable explicative quantitative
- Variable explicative qualitative (+ de 2 modalités)

5 Validation du modèle

6 Exemple

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients
- Rappel sur l'odds-ratio
- Lien entre OR, modèle Logit et coefficient de la régression
- Variable explicative quantitative
- Variable explicative qualitative (+ de 2 modalités)

5 Validation du modèle

6 Exemple

Interprétation des coefficients

Rappel sur l'odds-ratio

	M	$\bar{M}$
E^+	a	b
E^-	c	d

Odds-ratio : mesure d'association entre exposition et maladie

$$\text{OR} = \frac{ad}{bc} = \frac{\frac{a}{b}}{\frac{c}{d}}$$

$\left.\begin{array}{l} a = \mathbb{P}(M/E^+) \\ b = \mathbb{P}(\bar{M}/E^+) \end{array}\right\} \frac{a}{b}$: rapport de côtes (odd) d'être exposé

$\left.\begin{array}{l} c = \mathbb{P}(M/E^-) \\ d = \mathbb{P}(\bar{M}/E^-) \end{array}\right\} \frac{c}{d}$: rapport de côtes (odd) d'être non-exposé

Mesure de l'association entre maladie et exposition ? $\Rightarrow$ Odds-ratio

$$OR = \left(\frac{\mathbb{P}(M/E^+)}{\mathbb{P}(\bar{M}/E^+)} \Big/ \frac{\mathbb{P}(M/E^-)}{\mathbb{P}(\bar{M}/E^-)} \right)$$

Interprétation des coefficients

Remarque : si $\mathbb{P}(M)$ est faible ($< 10\%$) alors OR $\approx$ RR ($RR = \frac{P(M/E^+)}{P(M/E^-)}$)

$$OR = \frac{\mathbb{P}(M/E^+)}{\mathbb{P}(\bar{M}/E^+)} \times \frac{\mathbb{P}(\bar{M}/E^-)}{\mathbb{P}(M/E^-)} = \underbrace{\frac{\mathbb{P}(M/E^+)}{\mathbb{P}(M/E^-)}}_{RR} \times \underbrace{\frac{\mathbb{P}(\bar{M}/E^-)}{\mathbb{P}(\bar{M}/E^+)}}_{\approx 1}$$

Interprétation de l'odds-ratio :

- $OR = 1$: pas d'association
- $OR > 1$: E^+ est un facteur de risque de M
- $OR < 1$: E^+ est un facteur protecteur de M

Remarque : si le test du χ^2 est non significatif alors $OR = 1$

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients
- Rappel sur l'odds-ratio
- Lien entre OR, modèle Logit et coefficient de la régression
- Variable explicative quantitative
- Variable explicative qualitative (+ de 2 modalités)

5 Validation du modèle

6 Exemple

Interprétation des coefficients

Lien entre OR, modèle Logit et coefficient de la régression :

Considérons une seule variable explicative X binaire $\begin{cases} 1 : E^+ \\ 0 : E^- \end{cases}$

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 X$$

$$\text{logit}(\pi(1)) = \beta_0 + \beta_1$$

$$\text{logit}(\pi(0)) = \beta_0$$

$$\text{logit}(\pi(1)) - \text{logit}(\pi(0)) = \beta_1$$

Interprétation des coefficients

Lien entre OR, modèle Logit et coefficient de la régression :

$$\text{logit}\left(\frac{\pi(1)}{\pi(0)}\right) = \log \underbrace{\left[\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}\right]}_{OR} = \beta_1$$

On en déduit que :

$$OR = e^{\beta_1}$$

L'exponentiel du coefficient peut être interprété comme un odds-ratio

Remarque : idem dans le cas multiple mais les OR sont ajustés sur les autres variables X_j

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients
- Rappel sur l'odds-ratio
- Lien entre OR, modèle Logit et coefficient de la régression
- Variable explicative quantitative
- Variable explicative qualitative (+ de 2 modalités)

5 Validation du modèle

6 Exemple

Interprétation des coefficients

Supposons que X soit quantitative :

$$OR = e^{\beta_1} = OR^{X=x_0+1/X=x_0} \ \forall x_0$$

$\Rightarrow$ OR quand X augmente d'une unité, quelque soit la valeur de X (x_0).

Exemple : X : age en dizaines d'années et $OR = 2$.

Passer de 20 à 30 ans multiplie par 2 le risque de maladie

$\equiv$

Passer de 60 à 70 ans multiplie par 2 le risque de maladie

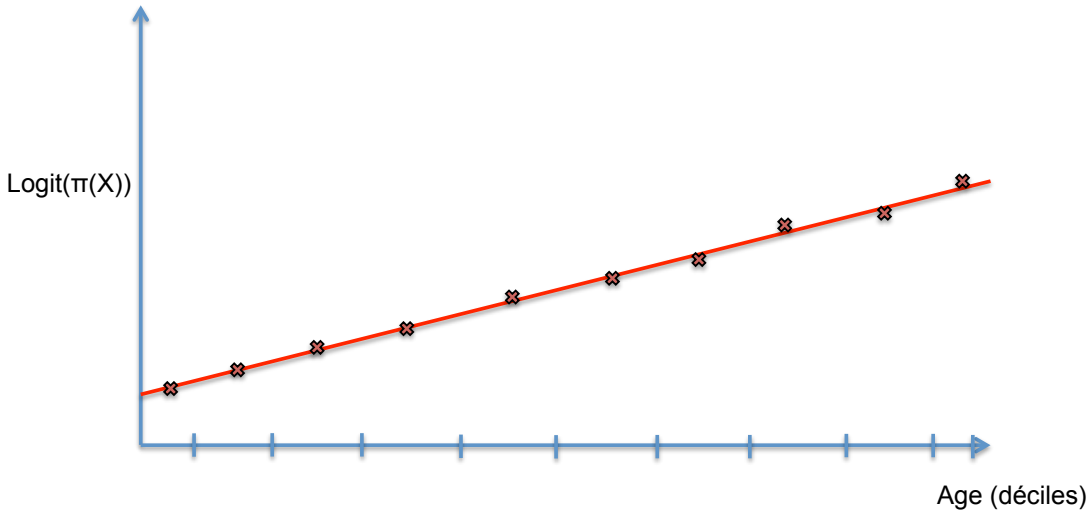
Cela sous-tend une hypothèse forte : log-linéarité de X qui est à vérifier.

Interprétation des coefficients

Principe

- Découper X en déciles
- Pour chaque intervalle on calcule $\mathbb{P}(Y = 1/X = c_1)$ (proportion de malades)
- Représenter graphiquement Logit$(\pi(X))$ en fonction des déciles de X

Objectif : vérification de la présence d'une relation linéaire entre X et Logit$(\pi(X))$



Sinon :

- Transformations mathématiques (log(X), $\sqrt{X}$,...)
- Discrétisation de X en classes appropriées

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients
- Rappel sur l'odds-ratio
- Lien entre OR, modèle Logit et coefficient de la régression
- Variable explicative quantitative
- Variable explicative qualitative (+ de 2 modalités)

5 Validation du modèle

6 Exemple

Interprétation des coefficients

Cas des variables nominales :

Exemple : niveau de conscience $\begin{cases} \text{normal} \\ \text{Coma leger} \\ \text{Coma profond} \end{cases}$

1. On choisit une modalité de référence (normal)
2. On construit 2 variables binaires $\begin{cases} \text{Coma leger}(0/1) \\ \text{Coma profond}(0/1) \end{cases}$
3. Introduction dans le modèle
   - Test de la variable dans sa totalité ($\{CL \bigcup CP\}$) (TRV)
   - Test des variables binaires une par une (test individuel)

Note : géré par les logiciels.

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle
- Pouvoir discriminant du modèle
- Calibration du modèle

6 Exemple

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle
- Pouvoir discriminant du modèle
- Calibration du modèle

6 Exemple

Pouvoir discriminant du modèle

Pouvoir discriminant

Capacité du modèle à correctement classer les observations (ex M / $\bar{M}$)

- Basé sur la courbe ROC (outil graphique d'évaluation du pouvoir discriminant)
- Critère *AUC* (*Area Under Curve*) = pouvoir discriminant du modèle

AUC	Discrimination
0.5	Nulle
0.7 - 0.8	Acceptable
0.8 - 0.9	Excellente
> 0.9	Exceptionnelle

Remarques

- Si $AUC = 0.5$ alors le modèle classe de manière complètement aléatoire les observations
- Si $AUC > 0.9$ le modèle est très bon, voire trop bon, il faut évaluer s'il y a *overfitting*

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle
- Pouvoir discriminant du modèle
- Calibration du modèle

6 Exemple

Calibration du modèle

Calibration

Comparaison des probabilités prédites par le modèle $\hat{\pi}_i(X_j)$ à celles observées dans l'échantillon. $\Rightarrow$ Mesure d'adéquation

Idée

On cherche à avoir un modèle qui minimise la distance entre les probabilités observées et celles prédites par le modèle

Détermination de la calibration $\Rightarrow$ Test de *Hosmer - Lemeshow*

Test de Hosmer - Lemeshow

Principe

On calcule pour chaque observation la probabilité prédite par le modèle $\hat{\pi}_i(X_j)$. On classe les observations par déciles de probabilités prédites. On compare dans chaque classe les effectifs observés et les effectifs théoriques.

- Si dans chaque classe ces deux effectifs sont proches alors le modèle est calibré
- S'il existe des classes dans lesquelles les effectifs sont trop différents, alors le modèle est mal calibré

Construction

1. Calculer les $\hat{\pi}_i(X_j)$ prédites par le modèle
2. Classer les données (observations + $\hat{\pi}_i(X_j)$) par ordre croissant de $\hat{\pi}_i(X_j)$
3. Regrouper les données par déciles de $\hat{\pi}_i(X_j)$
4. Construire le tableau suivant

Test de Hosmer - Lemeshow

	Malade (Y=1)		**Non-Malade (Y=0)**	
	Observés	*Prédits*	*Observés*	*Prédits*
G1 : 0 - 10%	#M	#prédits	#NM	#G1 - #prédits
G2 : 10% - 20%	.	.	.	.
G3 : 20% à 30%	.	.	.	.
G4 : 30 à 40%	.	.	.	.
G5 : 40% à 50%	.	.	.	.
G6 : 50% à 60%	.	.	.	.
G7 : 60% à 70%	.	.	.	.
G8 : 70 à 80%	.	.	.	.
G9 : 80% à 90%	.	.	.	.
G10 : 90 à 100%	.	.	.	.

- #M : le nombre de malades dans la classe (#NM : nb de non-malades)
- #prédits $= \sum_{G1} \hat{\pi}_i(X_j)$ car $Y = \pi(X) + \epsilon$

Test de Hosmer - Lemeshow

Hypothèses du test

- $\mathcal{H}_0$: les probabilités théoriques sont proches de celles observées (modèle calibré)
- $\mathcal{H}_1$: les probabilités théoriques sont différentes des observées (modèle non calibré)

Statistique de test

Sous $\mathcal{H}_0$

$$\widehat{C} = \underbrace{\sum_{G} \frac{(\#M - \#predits)^2}{\#predits}}_{\text{Malades}} + \underbrace{\sum_{G} \frac{(\#NM - \#predits)^2}{\#predits}}_{\text{Non Malades}} \sim \chi^2_{G-2} \text{ ddl}$$

- Le modèle est calibré si **on ne rejette pas** $\mathcal{H}_0$
- En pratique on ne rejette pas $\mathcal{H}_0$ si $p > 0.2$

Point étudié

1 Contexte général

2 Introduction à la régression logistique

3 Estimation des coefficients et tests

4 Interprétation des coefficients

5 Validation du modèle

6 Exemple

Exemple d'application

Etude sur les facteurs prénataux liés à un accouchement prématuré chez les femmes déjà en travail prématuré.

9 variables explicatives dans l'étude pour 390 femmes incluses dans l'étude.

Variable à expliquer : PREMATURE : accouchement prématuré (1=oui ; 0 = non)

Objectif : Quels sont les facteurs prédictifs d'un accouchement prématuré ??

`Source:http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`

Exemple d'application

Variables explicatives

- **GEST :** l'âge gestationnel en semaines à l'entrée dans l'étude
- **DILATE :** la dilatation du col en cm
- **EFFACE :** l'effacement du col (en %)
- **MEMBRAN :** les membranes rupturées (=1) ou non (=2) ou incertain (=3)
- **GRAVID :** la gestité (nombre de grossesses antérieures y compris celle en cours)
- **PARIT :** la parité (nombre de grossesses à terme antérieures)
- **DIAB :** la présence (=1) ou non (=2) d'un problème de diabète
- **TRANSF :** le transfert (1) ou non (2) vers un hôpital en soins spécialisés
- **GEMEL :** grossesse simple (=1) ou multiple (=2)

Code SAS :

```
proc logistic data=premature;
class membran(ref='2') diab(ref='2') transf(ref='2')
gemel(ref='1')/param=ref;
model premature(evt='1') = GEST DILATE EFFACE MEMBRAN
GRAVID PARIT DIAB TRANSF GEMEL / lackfit;
run;
```

Exemple d'application

Model Information	
Data Set	WORK.PREMATURE
Response Variable	PREMATURE
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	390
Number of Observations Used	390

Response Profile		
Ordered Value	**PREMATURE**	**Total Frequency**
1	0	124
2	1	266

Probability modeled is PREMATURE=1.

Exemple d'application

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	489.745	377.710
SC	493.711	421.338
-2 Log L	487.745	355.710

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	132.0345	10	<.0001
Score	103.6613	10	<.0001
Wald	71.8505	10	<.0001

Exemple d'application

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GEST	1	5.7887	0.0161
DILATE	1	9.6336	0.0019
EFFACE	1	12.1470	0.0005
MEMBRAN	2	24.5065	<.0001
GRAVID	1	2.3602	0.1245
PARIT	1	12.0120	0.0005
DIAB	1	2.3590	0.1246
TRANSF	1	4.3425	0.0372
GEMEL	1	3.4059	0.0650

Exemple d'application

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GEST	0.900	0.826	0.981
DILATE	1.632	1.198	2.223
EFFACE	1.017	1.007	1.026
MEMBRAN 1 vs 2	11.023	4.223	28.768
MEMBRAN 3 vs 2	2.145	0.514	8.950
GRAVID	1.243	0.942	1.641
PARIT	0.497	0.335	0.738
DIAB 1 vs 2	4.175	0.674	25.860
TRANSF 1 vs 2	1.791	1.035	3.099
GEMEL 2 vs 1	3.114	0.932	10.404

Exemple d'application

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.1	**Somers' D**	0.663
Percent Discordant	16.8	**Gamma**	0.664
Percent Tied	0.2	**Tau-a**	0.288
Pairs	32984	**c**	0.832

Exemple d'application

Partition for the Hosmer and Lemeshow Test					
		PREMATURE = 1		PREMATURE = 0	
Group	Total	Observed	Expected	Observed	Expected
1	39	12	8.80	27	30.20
2	40	13	14.89	27	25.11
3	39	21	18.30	18	20.70
4	39	17	22.25	22	16.75
5	39	26	26.36	13	12.64
6	39	30	29.91	9	9.09
7	39	33	33.65	6	5.35
8	39	38	36.35	1	2.65
9	39	39	37.81	0	1.19
10	38	37	37.68	1	0.32

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.3867	8	0.3107

Annexe 1 - Logarithme de la fonction de vraisemblance

$$\log(L) = \log\left[\prod_{i=1}^{n} \mathbb{P}(Y_i = y_i)\right] = \log\left[\prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}\right]$$

$$\log(L) = \log\left[\prod_{i=1}^{n} \pi(x_i)^{y_i}\right] + \log\left[\prod_{i=1}^{n} \left[1-\pi(x_i)\right]^{1-y_i}\right]$$

$$\log(L) = \sum_{i=1}^{n} y_i \log(\pi(x_i)) + \sum_{i=1}^{n}(1-y_i)\log(1-\pi(x_i))$$

$$\log(L) = \sum_{i=1}^{n} y_i \log\left(\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right) + \sum_{i=1}^{n}(1-y_i)\log\left(1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right)$$

Retour