Cours de Scoring - Séance 3 (Analyse du modèle de régression logistique)

Ibrahim TOURE,
Ingénieur Statisticien

Université d'Evry Val d'Essone

8 Décembre 2016

Sommaire

Notations

- On cherche à expliquer une variable Y binaire par p variables explicatives $X_1, \ldots, X_p$.
- On dispose de n observations $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in \mathbb{R}^p$, $y_i \in \{0, 1\}$.
- On note $\mathbb{X}$ la matrice $n \times p$ contenant les observations des variables explicatives :

$$\mathbb{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}.$$

Le modèle de régression logistique

- On suppose que les variables explicatives sont déterministes.

Modèle logistique, [Hosmer and Lemeshow, 2000]

Les observations y_i sont des réalisations de variables aléatoires Y_i indépendantes de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1-p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- **Remarque :** la variable X_1 peut correspondre à la constante du modèle. Dans ce cas, $x_{i1} = 1, i = 1, \ldots, n$.

Autre présentation du modèle logistique

- On suppose pour simplifier qu'on dispose d'une seule variable explicative X.
- On suppose qu'il existe une variable latente (inobservée) $Y^\star$

$$Y_i^\star = \tilde{\beta}_0 + \beta_1 x_i + \varepsilon$$

où ε est une variable aléatoire centrée, telle que

$$Y_i = \mathbf{1}_{Y_i^\star > s}, \quad s \in \mathbb{R}.$$

- On a alors

$$\mathbf{P}(Y_i = 1) = \mathbf{P}(-\varepsilon < \beta_0 + \beta_1 x_i) = F_\varepsilon(\beta_0 + \beta_1 x_i)$$

où $\beta_0 = \tilde{\beta}_0 - s$.

Figure :

Autre présentation du modèle logistique

- Définir le modèle revient à spécifier la loi de ε.

Propriété

- Si ε suit une loi logistique, c'est à dire de fonction de répartition

$$F_\varepsilon(x) = \frac{\exp(x)}{1+\exp(x)},$$

alors le modèle est le modèle logistique.

- Si ε suit une loi normale centrée réduite alors le modèle est le modèle probit.

Remarque

- Le choix de la fonction de lien dans le formalisme GLM correspond au choix de la loi de ε avec ce formalisme.
- Ce formalisme nous permettra d'introduire plus tard le modèle polytomique ordonné.

Figure :

Autre présentation du modèle logistique

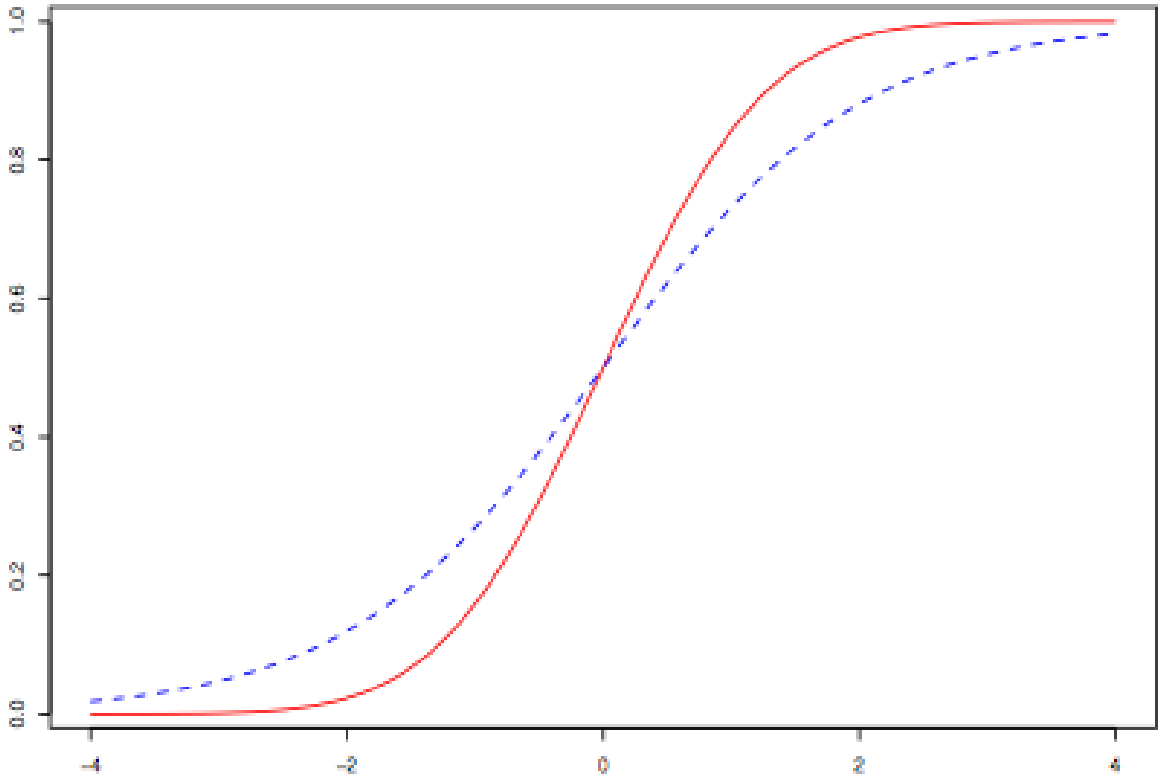


FIGURE: Fonctions de répartition pour le modèle logistique (bleu) et probit (rouge).

2 Types de données

- **Rappels :** on considère un n échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i \in \mathbb{R}^p$ sont déterministes et les Y_i sont des variables aléatoires de loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que

$$\text{logit}\, p_\beta(x_i) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta.$$

- On doit distinguer deux structures de données pour écrire les choses proprement (notamment la vraisemblance du modèle) :
  1. Données individuelles : tous les x_i sont différents. Dans ce cas les choses sont relativement simples puisque les Y_i suivent bien une loi de Bernoulli.
  2. Données répétées : il y a des répétitions sur les x_i. Il faut dans ce cas modifier légèrement les notations.

Figure :

Identifiabilité de la matrice design

- **Rappels :** la régression logistique modélise la loi de Y_i par une Bernoulli de paramètre $p_\beta(x_i)$.
- Par définition, le modèle est dit identifiable si $\beta \mapsto \mathbf{P}_{(Y_1,\dots,Y_n)}$ est injective.
- Le modèle logistique est donc identifiable si pour tout $\beta \neq \tilde{\beta}$ il existe $i \in \{1,\dots,n\}$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$.

Propriété

Si $n > p$ alors le modèle est identifiable si et seulement si $\text{rang}(\mathbb{X}) = p$.

Figure :

Rappels

- La matrice de design $\mathbb{X}$ contient "les observations des variables explicatives" :

$$\mathbb{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}.$$

- Elle joue un role important pour :
  1. l'identifiabilité du modèle ;
  2. l'estimation des paramètres du modèle ;
  3. le comportement asymptotique des estimateurs du modèle.

Figure :

Un exemple

- Un chef d'entreprise souhaite vérifier la qualité d'un type de machines en fonction de l'âge et de la marque des moteurs. Il dispose
  1. d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
  2. d'une variable quantitative `age` repésentant l'âge du moteur ;
  3. d'une variable qualitative a 3 modalités `marque` représentant la marque du moteur,

- et de $n = 33$ observations :

```
> panne
   etat age marque
1     0   4      A
2     0   2      C
3     0   3      C
4     0   9      B
5     0   7      B
```

Figure :

Variable quantitative

- C'est le cas le plus simple, un seul coefficient est dans le modèle par variable explicative $\Longrightarrow$ une variable quantitative est représentée par une seule colonne dans la matrice de design.

Exemple : pannes de machines

- On considère les modèles logistiques permettant d'expliquer `panne` par `age` et par `age` et la constante :

$$\text{logit}\, p_\beta(x_i) = \beta x_i \quad \text{et} \quad \text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 x_i.$$

- Les matrice de design associées à ces deux modèles sont

$$\mathbb{X} = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 9 \\ \vdots \end{pmatrix} \quad \text{et} \quad \mathbb{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 3 \\ 1 & 9 \\ \vdots & \vdots \end{pmatrix}.$$

Figure :

Variable qualitative

- Considérons le modèle logistique avec pour variable explicative `marque` pour l'exemple des **pannes de machines**. Une écriture naturelle est :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_i) + \beta_2 \mathbf{1}_B(x_i) + \beta_3 \mathbf{1}_C(x_i).$$

- Ce modèle n'est clairement pas identifiable.
- En effet, la matrice de design associée à ce modèle

$$\begin{bmatrix} A \\ C \\ C \\ B \\ B \\ \vdots \end{bmatrix} \Longrightarrow \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

n'est clairement pas de plein rang.

Figure :

Contraintes d'identifiabilité

- Le modèle précédent est surparamétré. Il est nécessaire de définir des contraintes d'identifiabilité.
- Les contraintes les plus utilisées sont $\beta_1 = 0$ (choix une modalité de référence) ou $\beta_1 + \beta_2 + \beta_3 = 0$.
- Le choix de la contrainte n'est pas forcément spécifié dans les sorties logiciels. Il est capital d'aller voir l'aide des fonctions pour connaitre les contraintes d'identifiabilité.

Remarque

Mathématiquement, le choix de la contrainte n'a pas une grande importance. Il doit en revanche être pris en compte pour pouvoir interpréter correctement les paramètres du modèle.

Figure :

Remarque

- Par défaut, R choisit comme modalité de référence la première modalité de la variable qualitative :

```
> glm(etat~marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC
     0.5596      -0.4261      -1.4759
```

- On peut modifier le choix de la modalité de référence

```
> glm(etat~C(marque,base=2),data=panne,family=binomial)

Coefficients:
          (Intercept)  C(marque, base = 2)1  C(marque, base = 2)3
               0.1335                0.4261               -1.0498
```

Figure :

Intéractions

Définition

Deux variables explicatives interagissent si l'effet de l'une de ces variables sur la variable à expliquer est différent selon les modalités de l'autre.

- On considère le modèle logistique permettant d'expliquer Y (`etat`) par X_1 (`marque`) et X_2 (`age`) :

$$\text{logit}\, p_\beta(x_i) = \beta_0 + \beta_1 \mathbf{1}_A(x_{i1}) + \beta_2 \mathbf{1}_B(x_{i1}) + \beta_3 \mathbf{1}_C(x_{i1}) + \beta_4 x_{i2},$$

  muni de la contrainte $\beta_1 = 0$.
- Ce modèle stipule que l'age de la machine agit linéairement sur $\text{logit}\, p_\beta(x_i)$ et que le coefficient de linéarité est le même pour toutes les marques.

Figure :

Intéractions

- Il est bien entendu possible d'envisager que l'effet de l'age sur l'état de la machine ne soit pas exactement le même pour toutes les marques :

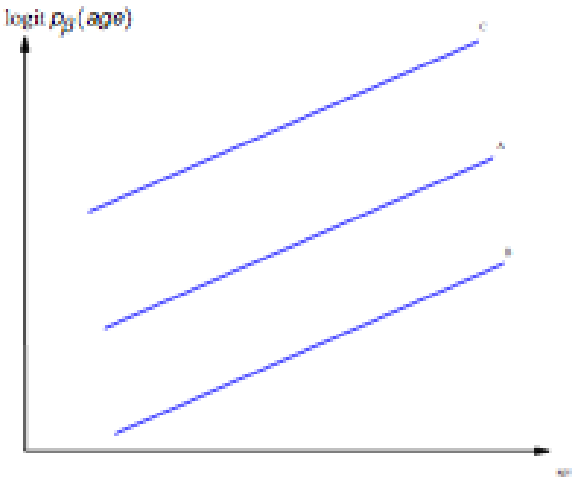


TABLE: Modèle additif.

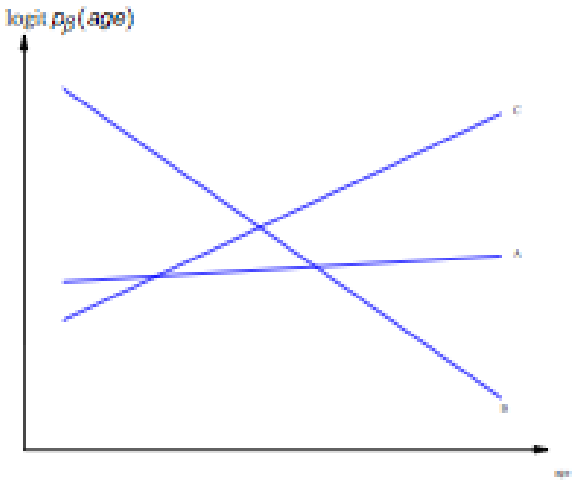


TABLE: Modèle avec interaction.

Intéractions

- La figure de droite correspond à un modèle du genre

$$\text{logit}(p_\beta(x_i)) = \begin{cases} \beta_{01} + \beta_{11} x_{i2} & \text{si } x_{i1} = A \\ \beta_{02} + \beta_{12} x_{i2} & \text{si } x_{i1} = B \\ \beta_{03} + \beta_{13} x_{i2} & \text{si } x_{i1} = C. \end{cases} \quad (1)$$

- Un tel modèle prend en compte l'interaction `marque:age` :

```
> glm(etat~marque+age+age:marque,data=panne,family=binomial)

Coefficients:
(Intercept)      marqueB      marqueC          age  marqueB:age  marq
    0.23512      0.19862     -2.43145      0.05641      -0.11188
```

- Le modèle ajusté ici n'est pas exactement celui défini par (1). Il est paramétré différemment :

$$\begin{aligned} \text{logit}(p_\beta(x_i)) = & \gamma_0 + \gamma_1 \mathbf{1}_A(x_{i1}) + \gamma_2 \mathbf{1}_B(x_{i1}) + \gamma_3 \mathbf{1}_C(x_{i1}) + \gamma_4 x_{i2} \\ & + \gamma_5 x_{i2} \mathbf{1}_A(x_{i1}) + \gamma_6 x_{i2} \mathbf{1}_B(x_{i1}) + \gamma_7 x_{i2} \mathbf{1}_C(x_{i1}) \end{aligned}$$

muni des contraintes $\gamma_1 = 0$ et $\gamma_5 = 0$.

Intéractions

- Il est facile de voir que les coefficients β_{jk} se déduisent des cofficients γ_j. Par exemple

$$\beta_{01} = \gamma_0 + \gamma_1, \quad \beta_{12} = \gamma_4 + \gamma_6, \quad \dots$$

- On peut bien évidemment ajuster directement le modèle (1) :

```
> glm(etat~marque-1+age:marque,data=panne,family=binomial)

Coefficients:
    marqueA      marqueB      marqueC  marqueA:age  marqueB:age  marq
    0.23512      0.43375     -2.19633      0.05641     -0.05547
```

- L'interaction présentée ci-dessus met en jeu une variable quantitative (`age`) avec une variable qualitative (`marque`). Il est bien entendu possible d'inclure des interactions entre variables qualitatives ou (plus rare) quantitatives.

Dimension d'un modèle

Définition

La dimension d'un modèle (logistique) est égale au nombre de paramètres identifiables du modèle. Elle correspond au nombre de colonnes de la matrice de design $\mathbf{X}$.

On calcule la dimension du modèle en sommant les contributions de chaque variables du modèle :

- si la constante est présente, elle a une contribution de 1 :
- une variable quantitative (non discrétisée) a une contribution de 1 ;
- une variable qualitative à K modalités a une contribution de $K - 1$;
- la contribution d'une interaction s'obtient en faisant le produit des contributions des variables qui interagissent.

Figure :

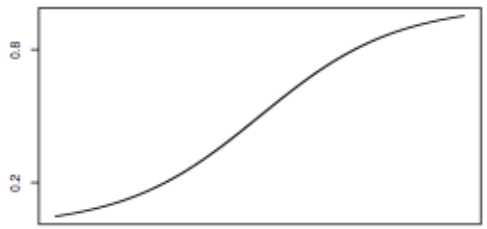
Interprétation des coefficients

- On considère l'allure de la courbe représentative de

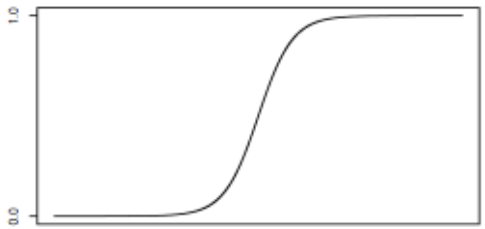
$$x \mapsto p_\beta(x) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}$$

pour $\beta = 0, 0.5, 2, 10$

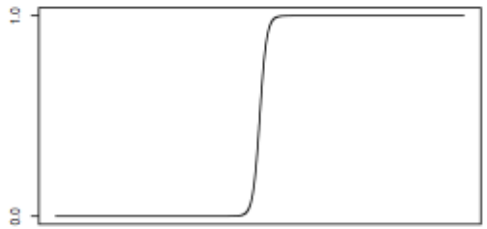
beta0



beta0.5



beta2



beta10

Lorsque β augmente, $p_\beta(x)$ est souvent proche de 0 ou 1.

Odds ratio

- On peut être tentés de dire : *plus β est grand, mieux on discrimine.*
- Prudence : tout dépend de l'échelle de x (si x change d'échelle, β va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

Définition

- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p_\beta(x)}{1 - p_\beta(x)}.$$

- L'odds ratio (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$OR(x, \tilde{x}) = \frac{\text{odds}(x)}{} = \frac{\frac{p_\beta(x)}{1-p_\beta(x)}}{}$$

Figure :

Interprétation des odds ratio

- Il faut être prudent avec l'interprétation des OR : ils sont très souvent utilisés mais pas toujours bien interprétés.

1. **Comparaison de probabilités de succès entre deux individus :**

$$\begin{array}{lll} OR(x,\tilde{x}) > 1 & \Longleftrightarrow & p_\beta(x) > p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) = 1 & \Longleftrightarrow & p_\beta(x) = p_\beta(\tilde{x}) \\ OR(x,\tilde{x}) < 1 & \Longleftrightarrow & p_\beta(x) < p_\beta(\tilde{x}) \end{array}$$

2. **Interprétation en termes de risque relatif :** dans le cas où $p(x)$ et $p(\tilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x,\tilde{x}) \approx p_\beta(x)/p_\beta(\tilde{x})$$

et interpréter "simplement".

Figure :

Interprétation des odds ratio

3. **Mesure de l'impact d'une variable :** pour le modèle logistique

$$\text{logit}\, p_\beta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

il est facile de vérifier que

$$\text{OR}(x, \tilde{x}) = \exp(\beta_1(x_1 - \tilde{x}_1)) \ldots \exp(\beta_p(x_p - \tilde{x}_p)).$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et $\tilde{x}$ qui diffèrent uniquement par la $j$ème variable. On obtient alors

$$\text{OR}(x, \tilde{x}) = \exp(\beta_j(x_j - \tilde{x}_j)).$$

Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

Sommaire

Problématique

- On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables $X_1, \ldots, X_p$ défini par

$$\text{logit}\, p_\beta(x) = \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

avec $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ qui contient les paramètres inconnus du modèle.

- Le caractère binaire de la variable à expliquer rend la méthode des moindres carrés impossible à mettre en oeuvre dans ce contexte.

- On rappelle que les estimateurs des moindres carrés du modèle linéaire gaussien coïncident avec les estimateurs du maximum de vraisemblance.

- C'est par cette approche que sont estimés les paramètres du modèle logistique à partir d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ (les variables aléatoires Y_i sont indépendantes).

La vraisemblance

- Les variables aléatoires $Y_1, \ldots, Y_n$ étant discrètes et indépendantes, la vraisemblance du modèle logistique est définie par

$$L_n : \{0,1\}^n \times \mathbb{R}^p \to \mathbb{R}$$

$$(y_1, \ldots, y_n, \beta) \mapsto \prod_{i=1}^{n} \mathbf{P}_\beta(Y_i = y_i)$$

où $\mathbf{P}_\beta$ désigne la probabilité sous le modèle logistique de paramètre β.

- Pour simplifier, on notera $L_n(y_1, \ldots, y_n, \beta) = L_n(\beta)$ et $\mathcal{L}_n(\beta) = \log(L_n(\beta))$.

Propriété

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\}.$$

Calcul du gradient

- Un moyen naturel de maximiser la log-vraisemblance est d'annuler son gradient

$$\nabla \mathcal{L}_n(\beta) = \left(\frac{\partial \mathcal{L}_n}{\partial \beta_1}(\beta), \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}(\beta) \right).$$

- On montre que

$$\nabla \mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left[x_i (y_i - p_\beta(x_i)) \right] = \mathbb{X}'(\mathbb{Y} - \mathbb{P}_\beta)$$

avec

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad \mathbb{P}_\beta = \begin{pmatrix} p_\beta(x_1) \\ \vdots \\ p_\beta(x_n) \end{pmatrix}.$$

Figure :

Equation du Score

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1,\ldots,p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Figure :

Existence et Unicité de l'EMV

Conséquence

Si il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0.$$

Cette équation est appelée équation du score.

- Résoudre les équations de score revient à résoudre p équations à p inconnues :

$$x_{1j}y_1 + \ldots + x_{nj}y_n = x_{1j}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + \ldots + x_{nj}\frac{\exp(x_n'\beta)}{1+\exp(x_n'\beta)},\ j = 1,\ldots,p$$

- Ce système n'est pas linéaire en β et n'admet pas de solutions explicites.
- Solution : utiliser des algorithmes itératifs qui convergent vers la solution d'où la nécessité d'étudier les propriétés analytiques de $\mathcal{L}_n(\beta)$.

Figure :

Existence et Unicité de l'EMV

Un premier résultat

Proposition

Soit $(x_i, y_i), i = 1, \ldots, n$ un nuage de points avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$. On suppose que la matrice de design $\mathbb{X}$ est de plein rang égal à p. Alors la log-vraisemblance

$$\mathbb{R}^p \to \mathbb{R} \tag{2}$$

$$\beta \mapsto \mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \log(1 + \exp(x_i' \beta)) \right\} \tag{3}$$

est strictement concave.

Conséquence importante

Un algorithme itératif convergera vers l'estimateur du maximum du vraisemblance lorsque celui-ci existe. Il n'y a pas de risque de tomber sur un maximum local.

Figure :

Existence et Unicité de l'EMV

Définition

On dit que l'estimateur du maximum de vraisemblance n'existe pas lorsque les équations de score n'admettent pas de solution finie.

- **Exemple :** On dispose d'un échantillon de taille $n = 200$:

	x_i	y_i
1	A	0
⋮	⋮	⋮
100	A	0
101	B	1
⋮	⋮	⋮
200	B	1

TABLE: Les 200 observations.

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x_i) = \beta_1 \mathbf{1}_{x_i=A} + \beta_2 \mathbf{1}_{x_i=B}$$

Existence et Unicité de l'EMV

- Il est facile de voir que lorsque $\beta_1 \to -\infty$ et $\beta_2 \to +\infty$, la vraisemblance $L_n(\beta)$ tend vers 1.
- Les équations de score n'admettent pas de solution finie et l'emv n'existe pas.

```
> n <- 100
> X <- factor(c(rep("A",n),rep("B",n)))
> X <- c(rep(0,n),rep(1,n))
> Y <- factor(c(rep(0,n),rep(1,n)))
> model <- glm(Y~X,family=binomial)
Message d'avis :
In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
```

- Une alerte nous prévient que l'algorithme permettant d'estimer les paramètres n'a pas convergé. On peut vérifier cette convergence avec la commande :

```
> model$converged
[1] FALSE
```

Existence et Unicité de l'EMV

Autre Exemple

On considère 2 jeux de données générés selon le protocole suivant :

1. Le premier est tel que
   - pour $i = 1, \ldots, 50$, x_i est le réalisation d'une loi uniforme sur $[-1, 0]$ et $y_i = 0$;
   - pour $i = 51, \ldots, 100$, x_i est le réalisation d'une loi uniforme sur $[0, 1]$ et $y_i = 1$.
2. Le second nuage correspond au premier nuage dans lequel on a posé $y_1 = 1$.

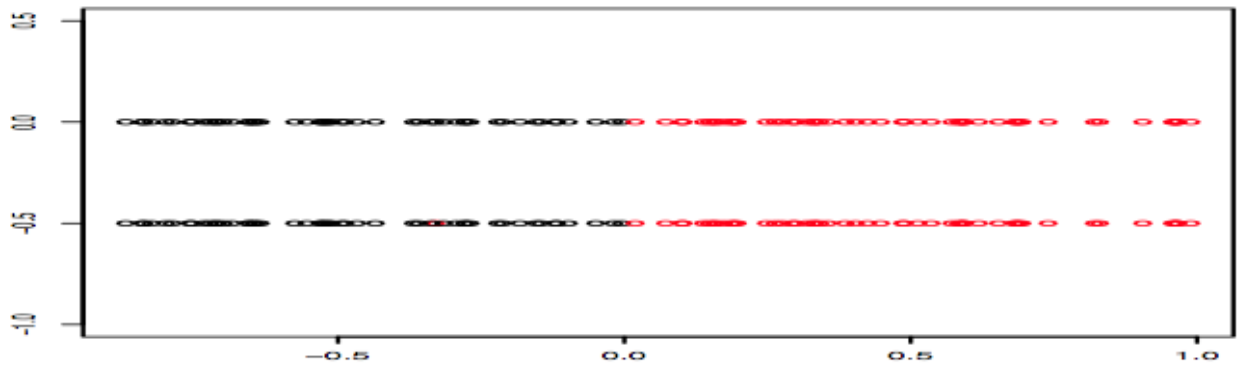


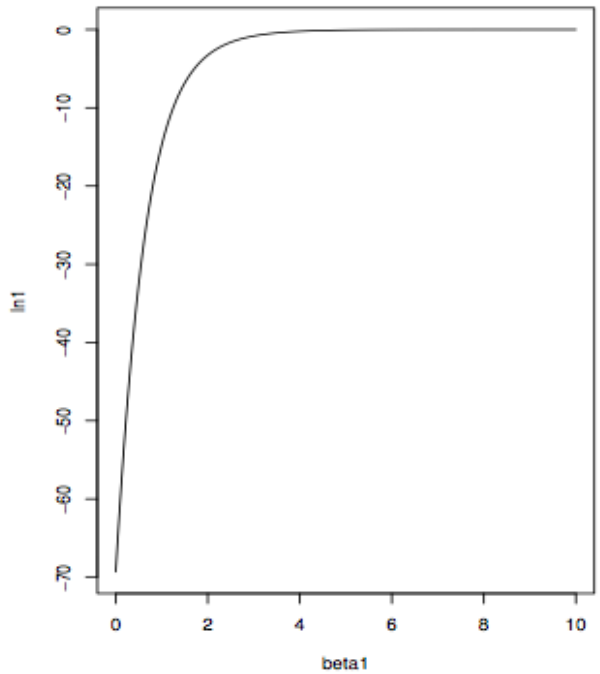
Figure :

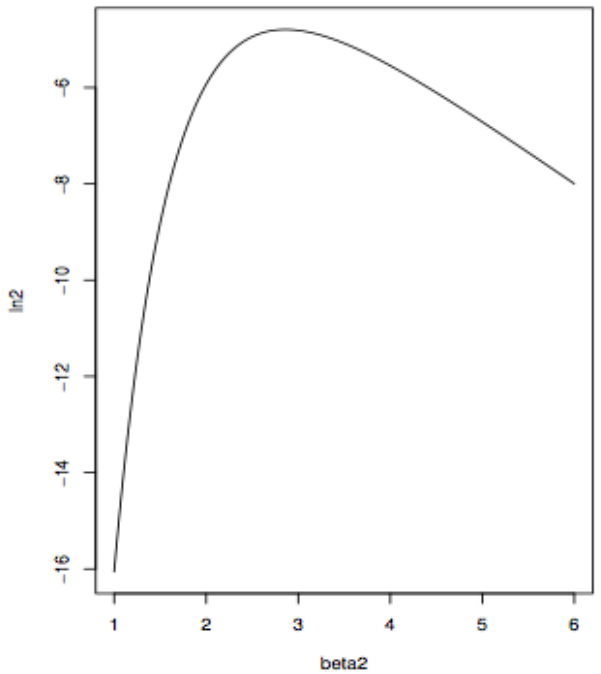
Existence et Unicité de l'EMV

- On considère le modèle logistique

$$\text{logit}\, p_\beta(x_i) = \beta x_i.$$

- La figure suivante représente $\beta \mapsto \mathcal{L}_n(\beta)$ pour les deux jeux de données.





Commentaires

- Pour les données 1, on voit que $\mathcal{L}_n(\beta) \to 0$ lorsque $\beta \to \infty \Longrightarrow$ l'emv n'existe pas.
- Pour les données 2, la vraisemblance admet un maximum unique.
- R nous prévient qu'il y a un problème pour le premier jeu de données :

```
> model1 <- glm(Y~X-1,data=nuage1,family=binomial)
Messages d'avis :
1: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  l'algorithme n'a pas convergé
2: In glm.fit(x = X, y = Y, weights = weights, start = start,
etastart = etastart,  :
  des probabilités ont été ajustées numériquement à 0 ou 1
> model1$converged
```

- Pour le second, tout se passe bien...

```
> model2 <- glm(Y~X-1,data=nuage2,family=binomial)
> model2$converged
[1] TRUE
```

Figure :

Commentaires

- Les cas où l'estimation ne se passe pas bien ont une caractéristique commune : les modalités de Y sont parfaitement séparées selon les valeurs de X.
- Les problèmes d'estimation interviennent dans des situations similaires à celles-ci.
- Albert et Anderson (1984) ont précisé cette notion de séparabilité.

Définition

Un nuage de points $(x_1, y_1), \ldots, (x_n, y_n)$ avec $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ est dit :

- complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta > 0$ et $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta < 0$;
- quasi-complètement séparable si $\exists \beta \in \mathbb{R}^p$: $\forall i$ tel que $Y_i = 1$ on a $x_i'\beta \geq 0$, $\forall i$ tel que $Y_i = 0$ on a $x_i'\beta \leq 0$ et $\{i : x_i'\beta = 0\} \neq \emptyset$;
- en recouvrement s'il n'est ni complètement séparable ni quasi-complètement séparable.

Commentaires

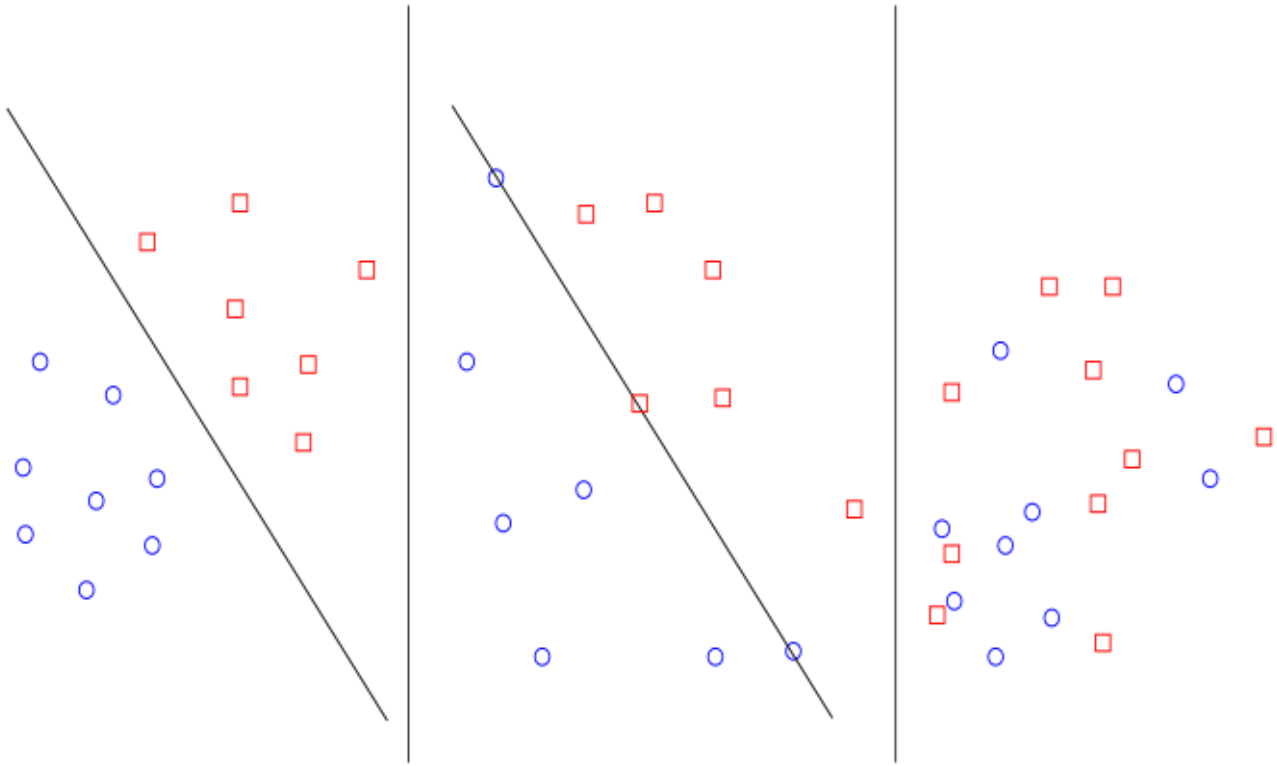


FIGURE: Exemple de séparabilité complète (gauche), quasi-complité (milieu) et de recouvrement (droite).

Existence de l'EMV

Très Important !

Théorème [Albert and Anderson, 1984]

- Si le nuage de points est complètement séparable ou quasi-complètement séparable alors l'estimateur du maximum de vraisemblance n'existe pas.
- Si le nuage de points est en recouvrement alors l'estimateur du maximum de vraisemblance existe et est unique.

- Il est important de réaliser que dans la plupart des cas réels, les données ne sont pas séparées.
- Par conséquent, dans la plupart des cas réels, l'emv existe et est unique.
- Nécessité de trouver des algorithmes itératifs qui vont converger vers l'emv.

Figure :

L'algorithme IRLS

- L'approche consiste à trouver une suite $(\beta^{(k)})_{k\in\mathbb{N}}$ de vecteurs de $\mathbb{R}^p$ qui converge vers l'estimateur du maximum de vraisemblance $\hat{\beta}_n$.
- On rappelle que, si il existe, $\hat{\beta}_n$ est solution de l'équation de score et vérifie donc

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = 0. \tag{4}$$

- Soit $\beta^{(k)}$ un vecteur de $\mathbb{R}^p$. Un développement de Taylor à l'ordre 1 donne l'approximation

$$S(\hat{\beta}_n) \approx S(\beta^{(k)}) + A(\beta^{(k)})(\hat{\beta}_n - \beta^{(k)}) \tag{5}$$

où $A(\beta^{(k)})$ désigne la matrice héssienne de la log-vraisemblance au point $\beta^{(k)}$:

$$A(\beta^{(k)}) = \nabla^2 \mathcal{L}_n(\beta^{(k)}) = -\mathbb{X}' W_{\beta^{(k)}} \mathbb{X}$$

- Ici $W_{\beta^{(k)}}$ désigne la matrice diagonale $n \times n$ de terme général

$$p_{\beta^{(k)}}(x_i)(1 - p_{\beta^{(k)}}(x_i)), \quad i = 1, \ldots, n.$$

L'algorithme IRLS

- Si $\mathbb{X}$ est de plein rang alors $A(\beta^{(k)})$ est inversible et on obtient en combinant (4) et (5)

$$\hat{\beta}_n \approx \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- Ce qui suggère d'utiliser la formule de récurrence

$$\beta^{(k+1)} = \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)}).$$

- D'où l'algorithme :

Algorithme IRLS

1. Initialisation : $\beta^{(0)}$, $k <- \quad 1$
2. Répéter jusqu'à convergence ($\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$)
   1. $\beta^{(k+1)} <- \quad \beta^{(k)} - A^{-1}(\beta^{(k)})S(\beta^{(k)})$.
   2. $k <- \quad k+1$

Pourquoi l'algorithme IRLS ?

- La formule de récurrence de l'algorithme de maximisation peut se réécrire

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbf{X}' (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} (\mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}}) \\ &= (\mathbb{X}' W_{\beta^{(k)}} \mathbb{X})^{-1} \mathbb{X}' W_{\beta^{(k)}} Z^{(k)},\end{aligned}$$

où $Z^{(k)} = \mathbb{X}\beta^{(k)} + {W_{\beta^{(k)}}}^{-1} (\mathbb{Y} - \mathbb{P}_{\beta^{(k)}})$.

- $\beta^{(k+1)}$ s'obtient en effectuant la régression pondérée du vecteur $Z^{(k)}$ par la matrice $\mathbb{X}$, d'où le nom de "Iterative Reweighted Least Square" (IRLS) pour cet algorithme.
- Les poids $W_{\beta^{(k)}}$ dépendent de $\mathbf{X}$ et $\beta^{(k)}$ et sont réévalués à chaque étape de l'algorithme.

Figure :

Comportement Asymptotique de l'emv

Propriétés générales de l'emv

- N'ayant pas d'écriture explicite pour l'emv, il est "difficile" d'étudier les propriétés de l'emv pour le modèle logistique (contrairement au modèle linéaire gaussien).
- Néanmoins on sait que, sous certaines hypothèses de régularité, l'emv $\hat{\theta}_n$ d'un paramètre θ vérifie certaines propriétés asymptotiques :
  1. Consistance : $\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta$ lorsque $n \to \infty$
  2. Normalité asymptotique :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}(\theta)^{-1})$$

où $\mathcal{I}(\theta)$ désigne la matrice d'information de Fisher du modèle au point θ.

Figure :

Comportement Asymptotique de l'emv

Théorème

Théorème [Fahrmeir and Kaufmann, 1985]

On suppose que :

- les $x_i, i = 1, \ldots, n$ prennent leurs valeurs dans une partie compacte de $\mathbb{R}^p$;
- la plus petite valeur propre $\lambda_{\min}(\mathbb{X}'\mathbb{X})$ tend vers $+\infty$ lorsque $n \to \infty$.

Alors

1. l'estimateur du maximum de vraisemblance existe asymptotiquement, c'est-à-dire qu'il existe une suite $\{\hat{\beta}_n\}_n$ de variables aléatoires

$$\lim_{n\to\infty} \mathbf{P}(S_n(\hat{\beta}_n) = 0) = 1.$$

2. la suite $\{\hat{\beta}_n\}_n$ est convergente : $\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta$.
3. $\{\hat{\beta}_n\}_n$ est asymptotiquement normal :

Figure :

Critique des hypothèses

Théorème

Le théorème précédent repose sous deux hypothèses.

1. La compacité de l'espace des régresseurs n'est pas une hypothèse restrictive (En pratique, les valeurs des variables explicatives varient le plus souvent dans une partie compacte de $\mathbb{R}^p$).

2. La seconde hypothèse implique que l'information (au sens de Fisher) sur le paramètre β augmente lorsque le nombre d'observations tend vers $+\infty$. Elle est nécessaire pour augmenter la précision (en terme de diminution de la variance) de l'estimateur $\hat{\beta}_n$ lorsque le nombre d'observations augmente avec n.

Figure :

Information de Fisher

- Le thérorème précédent n'est pas exploitable tel quel pour construire des intervalles de confiance ou des procédures de tests sur les paramètres du modèle (l'information de Fisher $\mathcal{I}(\beta)$ dépend du paramètre β qui est inconnu).
- On remarque que

$$(\hat{\beta}_n - \beta)' n\mathcal{I}(\beta)(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)' \mathcal{I}_n(\beta)(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

où $\mathcal{I}_n(\beta)$ désigne l'information de Fisher relative à l'ensemble des observations.

Propriété

$$\mathcal{I}_n(\beta) = -\mathbf{E}[\nabla^2 \mathcal{L}_n(\beta)] = \mathbb{X}' W_\beta \mathbb{X}.$$

Figure :

Loi des estimateurs

- On estime $\mathcal{I}_n(\beta)$ par

$$\hat{\Sigma} = \mathcal{I}_n(\hat{\beta}_n) = \mathbb{X}' W_{\hat{\beta}_n} \mathbb{X}.$$

- On déduit de la convergence en probabilité de $\hat{\beta}_n$ vers β et des opérations classiques sur la convergence en loi (Slutsky) que

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

Propriété

On désigne par $\hat{\sigma}_j^2$ je $j$ème terme de la diagonale de $\hat{\Sigma}^{-1}$. On a alors

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi^2_1 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Figure :

Intervalle de confiance et tests

1. Intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$IC_{1-\alpha}(\beta_j) = [\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j + u_{1-\alpha/2}\hat{\sigma}_j]$$

où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. Test (asymptotique) de nullité d'un paramètre au niveau α :
   - $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.
   - Sous H_0, $T = \hat{\beta}_j/\hat{\sigma}_j$ suit (approximativement) une loi $\mathcal{N}(0,1)$.
   - On rejette H_0 si $T_{obs} > u_{1-\alpha/2}$.

Figure :

Exemple

On reprend les données sur les pannes de machines.

```
> model <- glm(etat~.,data=panne,family=binomial)
```

- On obtient les tests de nullité des paramètres avec la fonction **summary** :

```
> summary(model)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.47808    0.83301   0.574    0.566
age          0.01388    0.09398   0.148    0.883
marqueB     -0.41941    0.81428  -0.515    0.607
marqueC     -1.45608    1.05358  -1.382    0.167
```

- Pour les intervalles de confiance, on utilise **confint** :

```
> confint(model)
                 2.5 %     97.5 %
(Intercept) -1.1418097 2.2222689
age         -0.1721209 0.2086368
marqueB     -2.0793170 1.1657128
marqueC     -3.7421379 0.5176220
```

Figure :

Test de nullité de q coefficients

- Tester la nullité d'un paramètre n'est pas suffisant.

  1. Comment tester la nullité de tous les paramètres (à l'exception de la constante) ? Equivalent du test de Fisher en régression linéaire.
  2. Comment tester l'effet d'une variable explicative qualitative ? Pour tester l'effet de la variable `marque`, on teste la nullité simultanée des coefficients du modèle associé à cette variable

Necessité de développer des procédures de tests permettant de tester des hypothèses du genre :

$$H_0 : \beta_1 = \ldots = \beta_q = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0.$$

Figure :

Test de Wald

- Il est basé sur le résultat :

$$(\hat{\beta}_n - \beta)' \hat{\Sigma} (\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2_p.$$

- On désigne par $\hat{\beta}_n^{(q)}$ les q premières composantes de $\hat{\beta}_n$ et $\hat{\Sigma}^{(q)}$ la matrice $q \times q$ comprenant les q premières lignes et colonnes de $\hat{\Sigma}$. On a alors :

$$(\hat{\beta}_n^{(q)} - \beta^{(q)})' \hat{\Sigma}^{(q)} (\hat{\beta}_n^{(q)} - \beta^{(q)}) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On déduit que sous H_0

$$\hat{\beta}^{(q)} \hat{\Sigma}^{(q)} \hat{\beta}^{(q)} \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Figure :

Test de déviance ou du rapport de vraisemblance

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir

$$\hat{\beta}_{H_0} \approx \hat{\beta}_n \quad \text{et} \quad \mathcal{L}_n(\hat{\beta}_{H_0}) \approx \mathcal{L}_n(\hat{\beta}_n).$$

- Plus précisément, on montre que sous H_0,

$$2(\mathcal{L}_n(\hat{\beta}_n) - \mathcal{L}_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \chi^2_q.$$

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Figure :

Test du Score

- Idée : on note $\hat{\beta}_{H_0}$ l'emv contraint sous H_0. Si H_0 est vraie, on doit avoir $S(\hat{\beta}_{H_0}) = \nabla \mathcal{L}_n(\hat{\beta}_0) \approx 0$.

- Plus précisément, on montre que sous H_0,

$$S(\hat{\beta}_{H_0})' \hat{\Sigma}_{H_0}^{-1} S(\hat{\beta}_{H_0}) \xrightarrow{\mathcal{L}} \chi^2_q,$$

où $\hat{\Sigma}_{H_0} = \mathbb{X} W_{\hat{\beta}_{H_0}} \mathbb{X}$.

- On rejette l'hypothèse nulle si la valeur observée de la statistique de test ci dessus est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Figure :

Récapitulatif

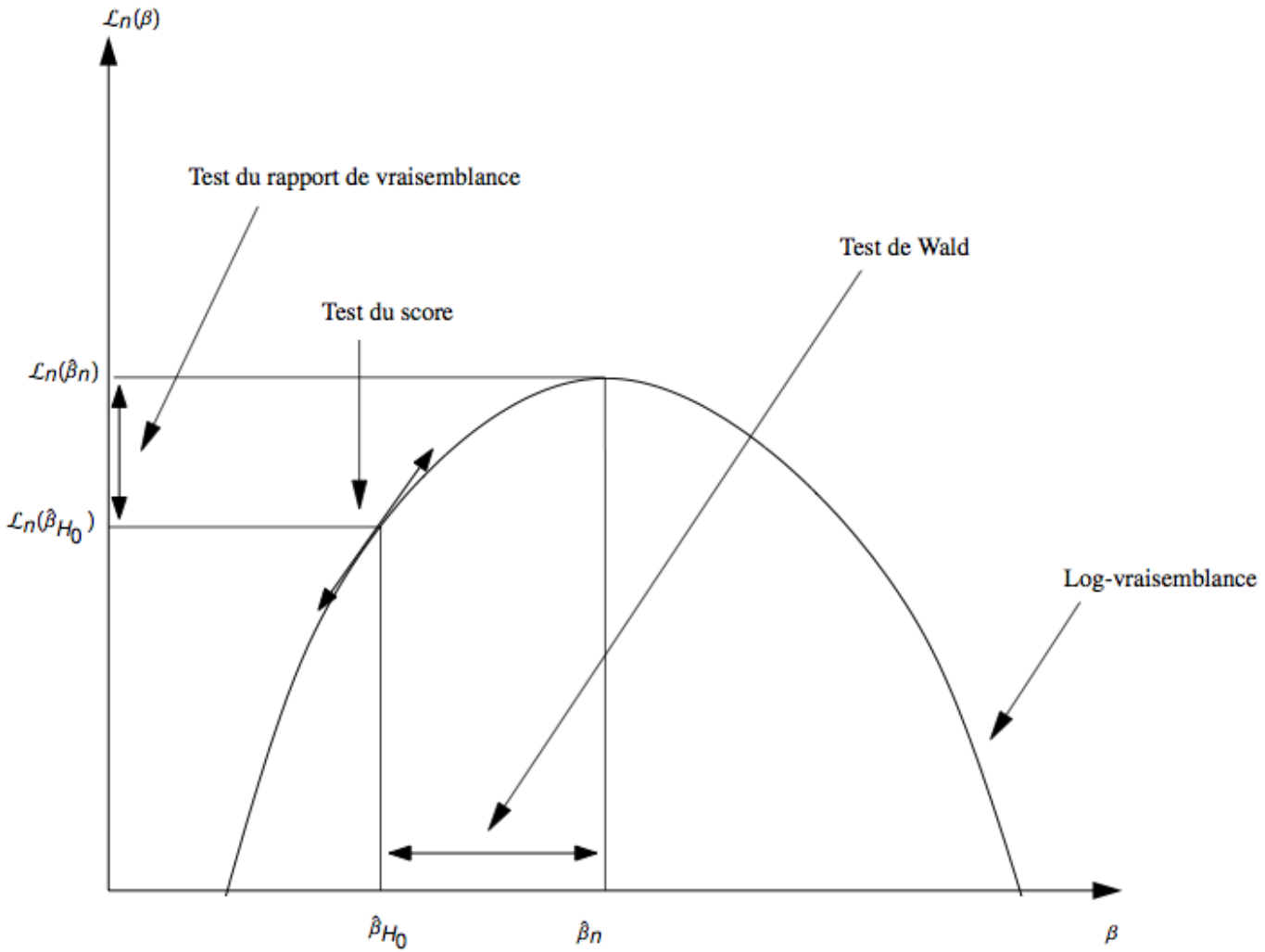


Figure :

Exemple sous R

- On peut tester l'effet des variables sour R avec la fonction **Anova** du package car :

1. Pour le test de Wald :

```
> library(car)
> Anova(model,type=3,test.statistic="Wald")
Analysis of Deviance Table (Type III tests)

Response: etat
             Df  Chisq Pr(>Chisq)
(Intercept)   1 0.3294     0.5660
age           1 0.0218     0.8826
marque        2 1.9307     0.3809
Residuals    29
```

2. Pour le test du rapport de vraisemblance :

```
> Anova(model,type=3,test.statistic="LR")
Analysis of Deviance Table (Type III tests)

Response: etat
       LR Chisq Df Pr(>Chisq)
age     0.02189  1     0.8824
marque  2.09562  2     0.3507
```

Figure :

Exemple sous SAS

- Sous SAS, on utilise la `proc logistic`

```
proc logistic data=Tp1_panne descending;
class marque;
model panne= age marque;
run;
```

Figure :

Exemple sous SAS

06:43 mardi, janvier 14, 2014

Le Système SAS

Procédure LOGISTIC

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et covariables
AIC	47.717	51.502
SC	49.214	57.488
-2 Log	45.717	43.502

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	2.2152	3	0.5290
Score	2.1630	3	0.5393
Wald	2.0333	3	0.5655

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
age	1	0.0218	0.8826
marque	2	1.9306	0.3809

Exemple sous SAS

06:50 mardi, janvier 14, 2014

Le Système SAS

Procédure LOGISTIC

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-0.1471	0.6265	0.0551	0.8144
age		1	0.0139	0.0940	0.0218	0.8826
marque	0	1	0.6252	0.5344	1.3684	0.2421
marque	1	1	0.2058	0.4907	0.1758	0.6750

Estimations des rapports de cotes			
Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
age	1.014	0.843	1.219
marque 0 vs 3	4.289	0.544	33.820
marque 1 vs 3	2.820	0.407	19.544

Figure :