MODELISATION DE DONNÉES QUALITATIVES

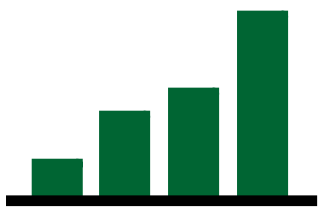
PREMIÈRE PARTIE

Pierre-Louis Gonzalez

I INTRODUCTION

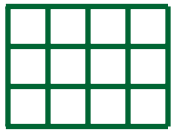
1 variable qualitative

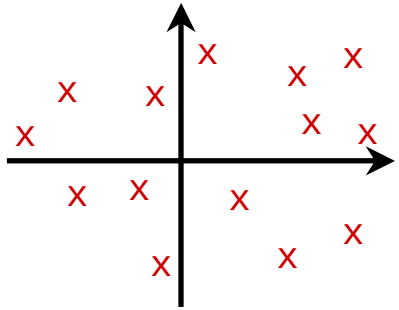
- Tri à plat
- Représentations graphiques



- Modélisation : loi binomiale
  loi multinomiale

2 variables qualitatives

- Tri croisé 
- Indépendance ?
- Khi-deux ...
- Description du tableau de contingence par analyse des correspondances simples.



Plus de deux variables qualitatives

- Tris croisés pour tous les couples de variables (tableau de Burt)

- Analyse des correspondances multiples

But de l'étude ?

- Modélisation

- Expliquer une variable à l'aide d'autres variables ...

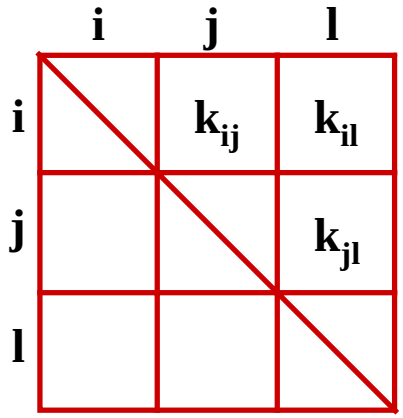
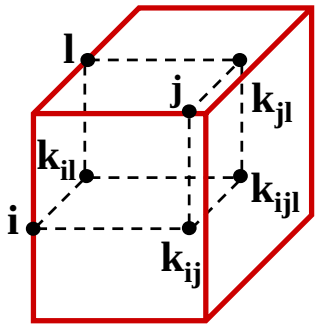
- **Effets de structure**

Le recours à l'utilisation de modèles (linéaires, logistiques) est nécessaire pour isoler les effets propres.

- Séparation des effets
- Effet d'une variable toutes choses égales par ailleurs
- Effet d'une variable conditionnellement aux variables introduites dans le modèle

Exemple **Vocations spécifiques de deux approches : description modélisation**

Correspondances multiples (DESCRIPTION)	Modèle log linéaire (EXPLORATION DE L'UNIVERS DES MODÈLES)
Description des liaisons entre les variables prises deux à deux sous forme essentiellement graphique.	Description des interactions entre plus de deux variables dans un cadre inférentiel.
N'impose aucune hypothèse sur les liaisons, mais impose une certaine homogénéité de l'ensemble des variables actives.	Des hypothèses sur les liaisons doivent être formulées au préalable.
N'est pas limitée dans le nombre de variables.	Est limité à peu de variables (en pratique moins de 5).

Correspondances multiples (DESCRIPTION)	**Modèle log linéaire (EXPLORATION DE L'UNIVERS DES MODELES)**
Met seulement en jeu les faces de l'hypercube représentées par le tableau de Burt.  	Met en jeu toutes les cases d'un hypercube de contingence. 
Les individus peuvent jouer un rôle central. L'analyse sert à produire des typologies d'individus.	Les individus n'apparaissent pas.

II LES MÉTHODES EXPLICATIVES

VARIABLE À EXPLIQUER	VARIABLES EXPLICATIVES $X_1, \ldots, X_K$		
Y	**Numériques**	**Nominales**	**Mixte**
Numérique	**Régression multiple** **REG** **GLM**	**Analyse de la variance** **GLM** **ANOVA**	**Analyse de la covariance** **GLM**
Qualitative	**Analyse discriminante** **CANDISC** **STEPDISC** **DISCRIM**	**DISQUAL** **Analyse discriminante sur variables qualitatives**	

VARIABLE À EXPLIQUER	VARIABLES EXPLICATIVES $X_1, \ldots, X_K$		
Y	**Numériques**	**Nominales**	**Mixte**
Nominale à deux modalités	**RÉGRESSION LOGISTIQUE** **LOGISTIC GENMOD**		
Nominale	**MODÈLE LINÉAIRE GÉNÉRALISÉ** **LOGISTIC CATMOD GENMOD**		
Ordinale	**RÉGRESSION LOGISTIQUE** **ou** **MODÈLE LINÉAIRE GÉNÉRALISÉ** **LOGISTIC** **CATMOD**		

III VARIABLE QUALITATIVE À EXPLIQUER

1 Variable dichotomique : $Y \in \{0,1\}$

Exemple 1 Soit P la population des ménages :

$$Y_i = \begin{cases} 1 & \text{si le ménage } i \in P \text{, possède un bien durable} \\ 0 & \text{sinon} \end{cases}$$

$X_i = (\text{AGE, CSP, SALAIRE, HABITAT, } \ldots)$ régresseurs

Exemple 2 Soit P la population des clients potentiels d'une banque :

«CREDIT SCORING»

$$\mathbf{Y_i} = \begin{cases} \mathbf{1} & \text{si un crédit est accordé au client } \mathbf{i} \\ \mathbf{0} & \text{sinon} \end{cases}$$

$$\mathbf{X_i} = (\text{AGE, REVENU, PRODUIT BANCAIRE, LIEU DE NAISSANCE, ...})$$

Exemple 3 Soit P la population des sujets testés à une dose «DOSAGE LEVEL»

$$\mathbf{Y_i} = \begin{cases} \mathbf{1} & \text{si le sujet } \mathbf{i} \in \text{P réagit au stimulus} \\ \mathbf{0} & \text{sinon} \end{cases}$$

$$\mathbf{X_i} = (\text{NIVEAU DE LA DOSE, POIDS, AGE, ...})$$

La variable réponse à expliquer $\mathbf{Y}$ est une variable de Bernoulli de paramètre $\mathbf{p_i}$.

$$\mathbf{p_i} = \mathbf{Pr}\left(\mathbf{Y_i} = \mathbf{1} \mid \mathbf{X_i}\right) = \mathbf{E}\left(\mathbf{Y_i} \mid \mathbf{X_i}\right)$$

$$Y_i \mid X_i \quad \rightarrow \quad B\left(1, p_i\right)$$

OBJECTIF

Exprimer $\mathbf{p_i}$ en fonction de $\mathbf{X_i}$

2 Variable polytomique

→ **Polytomique ordonnée**

Exemple 1 Soit P la population d'étudiants :

$$Y_i = \begin{cases} 1 & \text{si l'étudiant } i \in P \text{ pratique du sport tous les jours} \\ 2 & \text{si l'étudiant } i \in P \text{ pratique du sport une ou plusieurs fois par semaine} \\ 3 & \text{si l'étudiant } i \in P \text{ pratique du sport plus rarement} \end{cases}$$

La variable réponse $\mathbf{Y}$: «pratique du sport» est codée

$$\mathbf{X_i} = (\text{AGE, SEXE, TYPE D'ETUDES, } \ldots)$$

Exemple 2 Soit P la population de chômeurs à la date $\mathbf{t}$:

$$\mathbf{Y_i} = \begin{cases} \mathbf{1} & \text{si l'individu } \mathbf{i} \in \text{P est toujours au chômage à la date } \mathbf{t+\delta} \\ \mathbf{2} & \text{si l'individu } \mathbf{i} \in \text{P est en formation (stage)} \\ \mathbf{3} & \text{si l'individu } \mathbf{i} \in \text{P a un contrat CDD} \\ \mathbf{4} & \text{si l'individu } \mathbf{i} \in \text{P a un contrat CDI} \end{cases}$$

$$\mathbf{X_i} = (\text{AGE, SEXE, DIPLÔME, QUALIFICATION ...})$$

→ Polytomique non ordonnée

$\mathbf{Y_i}$ «distraction du samedi soir»

$$\mathbf{Y_i} = \begin{cases} \mathbf{1} = \text{télévision} \\ \mathbf{2} = \text{théâtre} \\ \mathbf{3} = \text{cinéma} \\ \mathbf{4} = \text{visite amis} \end{cases}$$

$$\mathbf{X_i} = (\text{AGE, SEXE, CSP, HABITAT, ...})$$

OBJECTIF

Exprimer $\mathbf{p_{ij} = P(Y_i = j \mid X_i)}$

en fonction de $\mathbf{X_i}$ pour $\mathbf{j = 1,\ 2,\ 3\ \ldots}$

IV POURQUOI DES MODÈLES PARTICULIERS ?

1 Cas de la régression linéaire classique

$$Y_i = x_i\beta + \varepsilon_i \qquad (1)$$

Y_i	$x_i\beta$	ε_i
variable aléatoire quantitative	prédicteur linéaire (élément déterminé)	variable aléatoire

- $\mathbf{E}\left(\varepsilon_i \mid \mathbf{X_i} = \mathbf{x_i}\right) = \mathbf{0}$

  Par la suite, on notera les espérances sans conditionnement $\mathbf{X_i} = \mathbf{x_i}$ ce qui revient à considérer $\mathbf{X_i}$ est non aléatoire. On notera indifféremment $\mathbf{X_i}$ ou $\mathbf{x_i}$.

- $\mathbf{V}\left(\varepsilon_i\right) = \sigma^2$

  Si de plus ε_i est supposée gaussienne, l'estimateur des moindres carrés ordinaire :

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

  est l'estimateur du maximum de vraisemblance.

2 Cas de la régression d'une variable dichotomique

Si on modélise par (1), on obtient un résidu qui est une v.a.r. discrète prenant deux valeurs :

$$(1) \Rightarrow \begin{cases} \varepsilon_i = 1 - x_i\beta & \text{avec la probalilité } p_i = P(Y_i = 1) \\ \varepsilon_i = -x_i\beta & \text{avec la probalilité } 1 - p_i \end{cases}$$

Si on modélise par (1) l'estimateur $\hat{\beta}$ n'est plus efficace.

$$(1) \Rightarrow E(Y_i) = x_i\beta$$

$$\text{Or } Y_i \rightarrow B(1, p_i) \Rightarrow E(Y_i) = p_i$$

$$\Rightarrow p_i = x_i\beta$$

T

une valeur qui n'est pas forcément entre 0 et 1

Le modèle (1) est donc inapproprié !

V NIVEAU D'UTILITÉ, VARIABLE LATENTE

1 Cas de variable latente

Z_i «intensité du désir de posséder le bien»
pour le ménage i caractérisé par x_i

Z_i non observable

$$\begin{cases} Y_i = 0 \Leftrightarrow Z_i < s \rightarrow (\text{seuil thérorique}) \\ Y_i = 1 \Leftrightarrow Z_i \geq s \end{cases}$$

c'est-à-dire $Y_i = \mathbb{1}_{(Z_i \geq s)}$

2 Fonction d'utilité

Soit $\mathbf{u(1, x_i)}$ le niveau d'utilité procuré par la possession du bien

$\mathbf{u(0, x_i)}$ le niveau d'utilité procuré par la non possession du bien

$$\begin{cases} \mathbf{Y_i = 0 \Leftrightarrow u(0, x_i) > u(1, x_i)} \\ \mathbf{Y_i = 1 \Leftrightarrow u(1, x_i) \geq u(0, x_i)} \end{cases}$$

c'est-à-dire :

$$\mathbf{Z_i = u(1, x_i) - u(0, x_i)}$$

$$\mathbf{Y_i = \mathbb{1}_{(Z_i \geq 0)}}$$

Dans ces deux cas, on peut exprimer la probabilité

$$\mathbf{p_i = P\left(Y_i = 1 \mid x_i\right)} \text{ comme :}$$

$$\mathbf{p_i = P\left(Z_i \geq s\right)}$$

VI MODÈLE THÉORIQUE

1 Données statistiques

$\left(Y_i, \quad x_i^1, \ \ldots \ , x_i^p\right)$ $\qquad$ $\mathbf{i = 1}$ à $\mathbf{n}$

variable dichotomique $\{0,1\}$

p variables explicatives quantitatives ou qualitatives

taille de l'échantillon

$\mathbf{X_i}$ vecteur de R^p

(On supposera $\left(x_i^1 = 1 \ \forall \ i \right)$

de façon à définir un modèle avec constante)

En introduisant Z variable latente non observable telle que :

$$\left.\begin{array}{l} Z_i = X_i\beta + \varepsilon_i \\ Y_i = \mathbb{1}_{Z_i > 0} \end{array}\right\} \quad p_i = P(Y_i = 1) = P(-\varepsilon_i < X_i\beta)$$

$$= F(X_i\beta)$$

fonction de répartition de $-\varepsilon_i$

2 Modèle stochastique général

$(\mathbf{Y_i, X_i})_{\mathbf{i=1\ldots n}}$ i.i.d tel que :

$$H_1 : Y_i \mid X_i \rightarrow B(1, p_i)$$

$$\mathbf{H_2 : p_i = P(Y_i = 1 \mid X_i) = F(X_i \beta)}$$

où $\mathbf{F : R \rightarrow [0,1]}$ fonction de répartition

Le paramètre β, vecteur de R^p formé des coefficients de régression est inconnu.

3 Modèles PROBIT, LOGIT, ...

L'hypothèse $\mathbf{H_2}$ dépend du choix de la fonction $\mathbf{F}$. Les modèles paramétriques usuels sont :

3.1 Le modèle probit

$$F(\omega) = \Phi(\omega) = \int_{-\infty}^{\omega} \frac{1}{\sqrt{2\Pi}} \exp\left(-\frac{t^2}{2}\right) dt \quad \forall\ \omega \in R$$

Fonction de répartition de la loi normale centrée réduite N(0;1)

$\mathbf{F^{-1} = \Phi^{-1}}$ **probit**

3.2 Le modèle logit

$$F(\omega) = \frac{e^{\omega}}{1+e^{\omega}} = \frac{1}{1+e^{-\omega}} \qquad \forall\ \omega \in \mathbb{R}$$

Fonction de répartition de la loi logistique

de moyenne **0** et de variance $\frac{\Pi^2}{3}$

$$F^{-1}(t) = \ln \frac{t}{1-t}$$ **logit**

3.3 Le modèle complémentaire log-log (ou modèle Gompit)

$$F(\omega) = 1 - \exp\left(-e^{\omega}\right) \quad \forall\ \omega \in \mathbb{R}$$

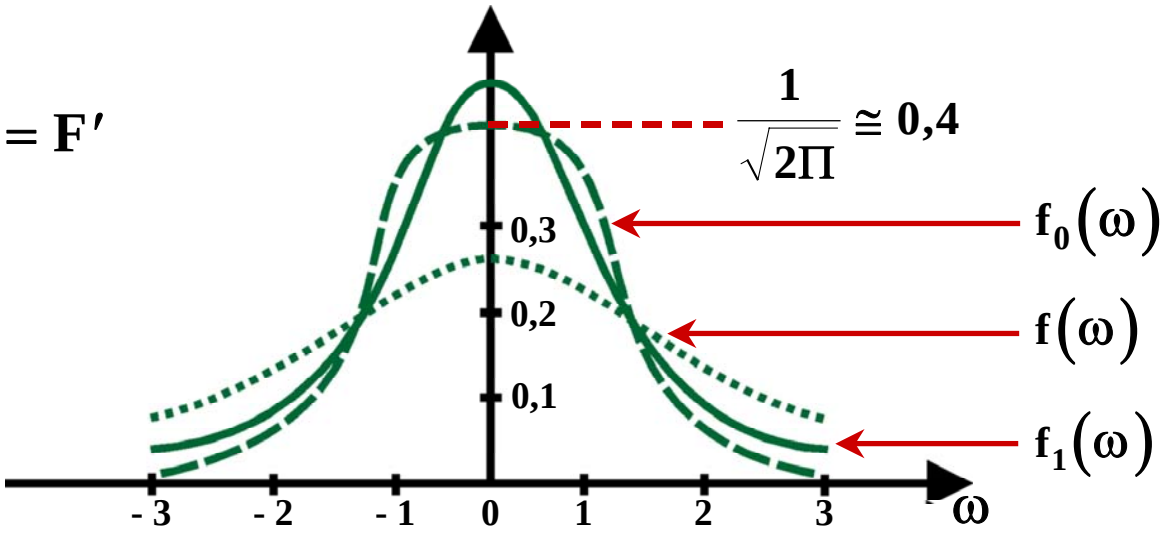
Fonction de répartition de la loi de Gompertz de moyenne 0,577 (constante d'Euler) et de variance $\frac{\Pi^2}{6}$

$$F^{-1}(t) = \ln\left(-\ln\left(1-t\right)\right)$$

Remarque Cette loi est dissymétrique.

4 Comparaison des modèles LOGIT et PROBIT

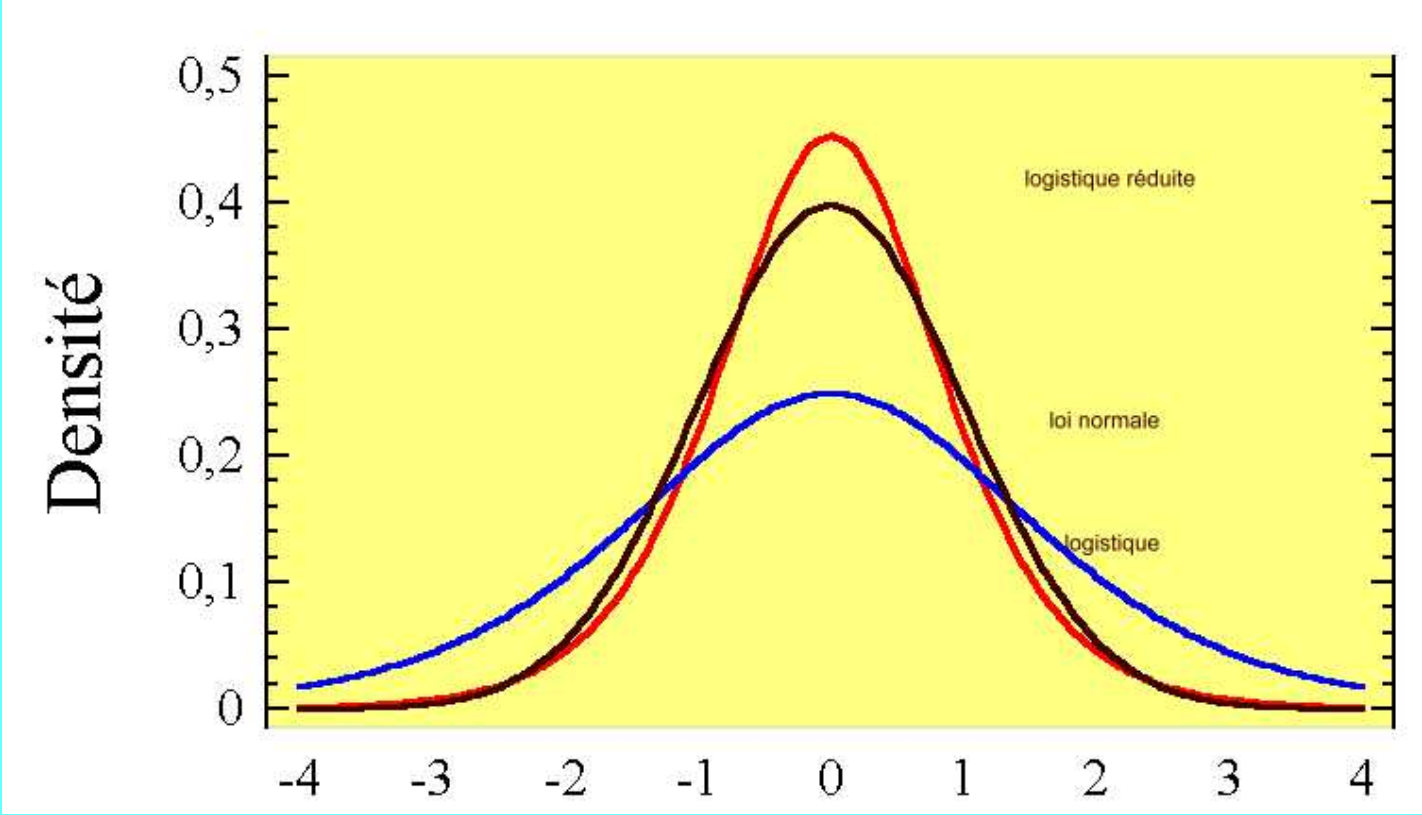
DENSITÉS $f = F'$



Modèle PROBIT $\Phi(\omega) = \int_{-\infty}^{\omega} \frac{1}{\sqrt{2\Pi}} e^{-t^2/2}\, dt \Rightarrow f_0(\omega) = \frac{1}{\sqrt{2\Pi}} e^{-\omega^2/2}$

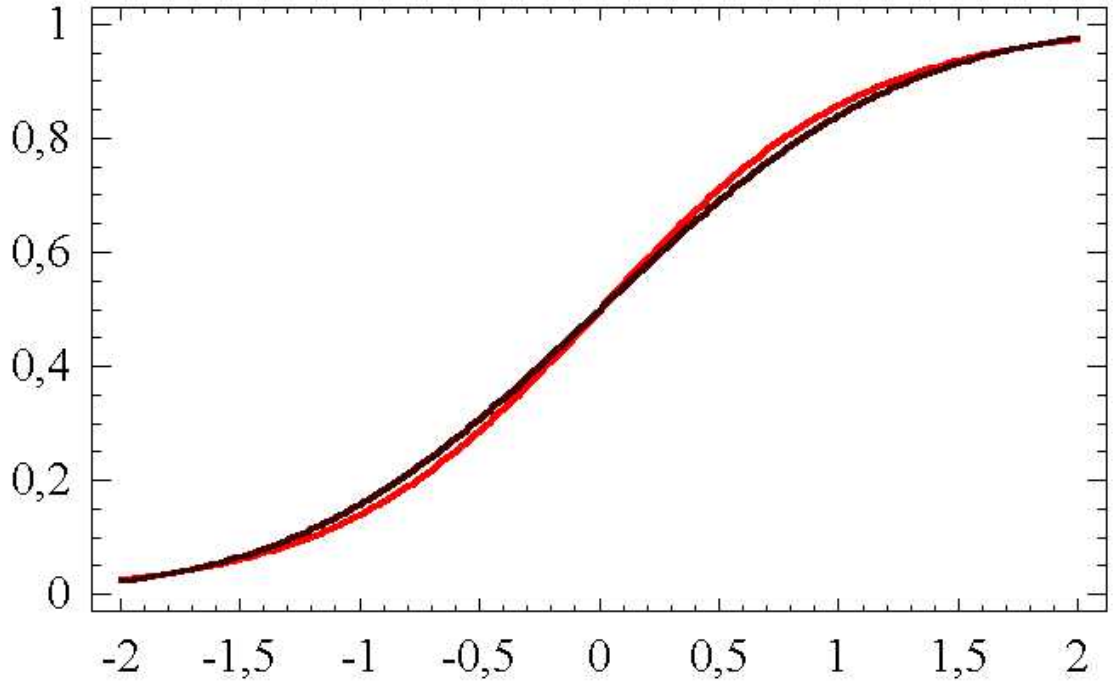
Modèle LOGIT $F(\omega) = \frac{1}{1+e^{-\omega}} \Rightarrow f(\omega) = \frac{e^{\omega}}{\left(1+e^{\omega}\right)^2}$

Modèle LOGIT réduit $F_1(\omega) = \frac{1}{1+e^{-\Pi\omega/\sqrt{3}}} \quad f_1(\omega) = \frac{\Pi}{\sqrt{3}} \frac{e^{\Pi\omega/\sqrt{3}}}{\left(1+e^{\Pi\omega/\sqrt{3}}\right)^2}$

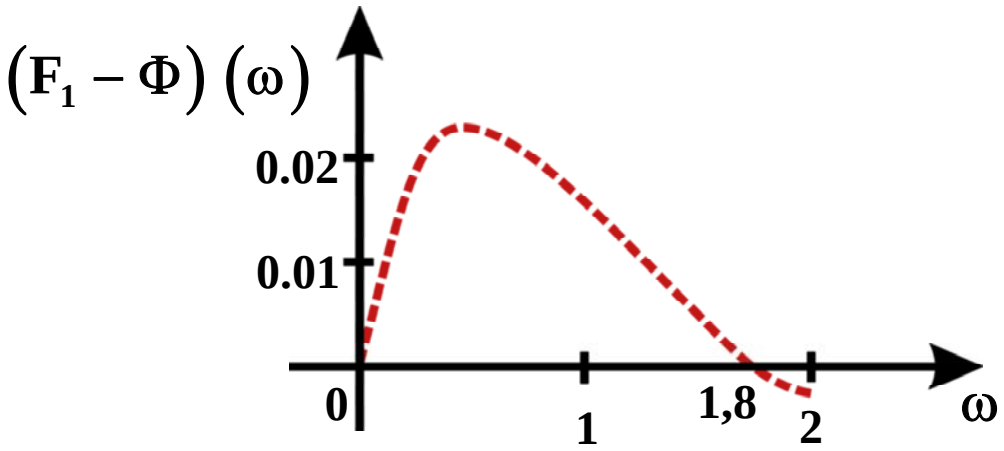


Sur les extrêmes la loi logistique s’approche un peu plus lentement de 0 ou de 1.

Comparaison des fonctions de répartition



DIFFÉRENCE



CONCLUSION

- Les lois F_1 et Φ sont proches $\Rightarrow$ modèles équivalents.

- La précédure de SAS, «LOGISTIC» utilise F ou Φ.

  $\Rightarrow$ Les estimateurs obtenus avec F (logit) seront $\Pi / \sqrt{3}$ fois plus grands qu'avec Φ (probit).

- Le modèle LOGIT est préférable car les calculs sont plus simples.

Dans la plupart des cas pratiques, on peut donc choisir indifféremment l'un ou l'autre modèle.

Le **modèle LOGIT** a l'avantage d'une plus grande simplicité numérique.

Le **modèle PROBIT** est en revanche plus proche du modèle habituel de régression par les moindres carrés.

Avantages du modèle LOGIT

Les coefficients du modèle LOGIT sont interprétables en termes d'odds-ratio.

Un échantillonnage ne respectant pas les proportions réelles dans la population des deux modalités de la variable à expliquer Y ne change que la constante dans le modèle.

VII PRINCIPES GÉNÉRAUX : ESTIMATION DU VECTEUR β

1 La méthode du maximum de vraisemblance

Échantillon $x_1 \dots x_n$

Loi de probabilité de x_i $f(x_i, \theta)$ où $\theta = (\theta_1, \dots, \theta_k)' \in \Omega$

Vraisemblance

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$ dépend des x_i et des θ

Estimation du maximum de vraisemblance

$$\hat{\theta} \qquad L(\hat{\theta}) = \text{MAX}_{\theta \in \Omega} \; L(\theta)$$

On obtient en général $\hat{\theta}$ en annulant les dérivées premières $\dfrac{\partial \, \text{Log} \, L(\theta)}{\partial \, \theta_i}$

Scores

$$\mathbf{u}_i(\theta) = \frac{\partial \, \mathbf{Log}\, \mathbf{L}(\theta)}{\partial \, \theta_i}$$

$$\mathbf{u}(\theta) = \left(\mathbf{u}_1(\theta), \ \ldots \ , \mathbf{u}_k(\theta)\right)' \ = \text{vecteur score}$$

On a : $\mathbf{u}\left(\hat{\theta}\right) = \mathbf{0}$

Matrice d'information de Fisher

$$\mathbf{I}(\theta) = \mathbf{E}\left[\frac{-\partial^2 \, \mathbf{Log}\, \mathbf{L}(\theta)}{\partial \, \theta^2}\right]$$

estimée par : $\hat{\mathbf{I}}\left(\hat{\theta}\right) = \left(\frac{-\partial^2 \, \mathbf{Log}\, \mathbf{L}(\theta)}{\partial \, \theta^2}\right)_{\theta=\hat{\theta}}$

Résultats

$\theta =$ vecteur des paramètres

$\hat{\theta} =$ estimation du maximum de vraisemblance

1. $\hat{\theta} \rightarrow N\left[\theta ; I(\theta)^{-1}\right]$

2. $u(\theta) \rightarrow N(0 ; I(\theta))$

3. $(\hat{\theta} - \theta)' \, \mathbf{I}(\theta)(\hat{\theta} - \theta) \rightarrow \chi^2_{(k)}$

4. $\mathbf{u}(\theta)' \, \mathbf{I}(\theta)^{-1} \mathbf{u}(\theta) \rightarrow \chi^2_{(k)}$

5. $\Lambda = -2 \operatorname{Log} \dfrac{\mathbf{L}(\theta)}{\mathbf{L}(\hat{\theta})} \rightarrow \chi^2_{(k)}$

2 Test global

$$\mathbf{H_0} : \theta = \theta_0$$

Statistiques

1. Statistique de Wald

$$\left(\hat{\theta} - \theta_0\right)' \mathbf{I}(\theta_0) \left(\hat{\theta} - \theta_0\right) \rightarrow \chi^2_{(k)} \text{ sous } \mathbf{H_0}$$

2. Statistique du score

$$u(\theta_0)' I(\theta_0)^{-1} u(\theta_0) \rightarrow \chi^2_{(k)} \text{ sous } \mathbf{H_0}$$ Avantage : pas de calcul de $\hat{\theta}$

3. Statistique des vraisemblances

$$\Lambda = -2 \text{ Log } \frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\hat{\theta})} \rightarrow \chi^2_{(k)} \quad \text{sous } \mathbf{H_0}$$

3 Test partiel

$\theta = (\theta_1, \theta_2)$ θ_1 a **p** coordonnées

$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) =$ estimation du M. V.

Test $\mathbf{H_0 : \theta_1 = \theta_{10}}$

On calcule $\hat{\theta}_{H_0} = (\theta_{10}, \tilde{\theta}_2)$

avec $\mathbf{L}(\theta_{10}, \tilde{\theta}_2) = \mathbf{\max_{\theta_2}} \mathbf{L}(\theta_{10}, \theta_2)$

$\hat{\theta}_{H_0} =$ estimation de θ sous $\mathbf{H_0}$

Statistiques utilisées

1. Wald $\left(\hat{\theta}_1 - \theta_{10}\right)' \mathrm{Var}\left(\hat{\theta}_1\right)^{-1}\left(\hat{\theta}_1 - \theta_{10}\right) \rightarrow \chi^2_{(p)}$ sous $\mathbf{H_0}$

$\mathbf{Var}\left(\hat{\theta}_1\right)$ est extrait de $\mathbf{Var}\left(\hat{\theta}\right) = \hat{\mathbf{I}}\left(\hat{\theta}\right)^{-1}$

2. Score

$$\mathbf{u}\left(\hat{\theta}_{H_0}\right)' \hat{\mathbf{I}}\left(\hat{\theta}_{H_0}\right)^{-1}\mathbf{u}\left(\hat{\theta}_{H_0}\right) \rightarrow \chi^2_{(p)} \quad \text{sous } \mathbf{H_0}$$

3. Rapport de vraisemblance

$$\Lambda = -2\,\mathbf{Log}\,\frac{\mathbf{L}\left(\theta_{10}, \tilde{\theta}_2\right)}{\mathbf{L}\left(\hat{\theta}_1, \hat{\theta}_2\right)} \rightarrow \chi^2_{(p)} \quad \text{sous } \mathbf{H_0}$$