

Économétrie des Variables Qualitatives

Modèles dichotomiques

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Cadre général – Variable dichotomique

- ▶ Une variable $y \in \{0, 1\}$ suit une **loi de Bernoulli** de paramètre p :

$$\mathbb{P}(y = k) = p^k(1 - p)^{1-k} = \begin{cases} p & \text{si } k = 1 \\ 1 - p & \text{si } k = 0 \end{cases}$$

- ▶ **Espérance** : $\mathbb{E}[y] = 0 \times (1 - p) + 1 \times p = p$.
- ▶ **Variance** : $\mathbb{V}[y] = p(1 - p)$. Maximale pour $p = 0.5$, s'annule quand $p \rightarrow 0$ ou $p \rightarrow 1$.

Cadre général – Modèle conditionnel

- ▶ On observe N variables de Bernoulli (y_1, \dots, y_N) **indépendantes** avec des paramètres **différents** (p_1, \dots, p_N) .
- ▶ Les probabilités dépendent des variables explicatives X_i :

$$p_i = p(X_i, \beta), \quad i = 1, \dots, N$$

où β est le paramètre à estimer.

- ▶ Contrairement au modèle marginal (statistique classique), chaque observation a sa propre probabilité \Rightarrow modèle **conditionnel**.

Cadre général – Log-vraisemblance

- ▶ **Log-vraisemblance :**

$$\ell(y | X, \beta) = \sum_{i=1}^N [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$$

où $p_i = p(X_i, \beta)$.

- ▶ **Score :**

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^N \frac{\partial p_i}{\partial \beta} \cdot \frac{y_i - p_i}{p_i(1 - p_i)}$$

- ▶ Cette représentation est valable pour **tous** les modèles dichotomiques. Il suffit de spécifier la forme de $p(X_i, \beta)$.

Cadre général – Modèle latent

- ▶ On suppose que la variable dichotomique observée résulte d'une variable **continue inobservable** y_i^* :

$$y_i^* = X_i\beta + u_i$$

où u_i est une perturbation d'espérance nulle.

- ▶ Règle de décision :

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \leq 0 \end{cases}$$

- ▶ Le seuil 0 est conventionnel (sans incidence si le modèle contient une constante).

Cadre général – Probabilité

- Calcul de la probabilité :

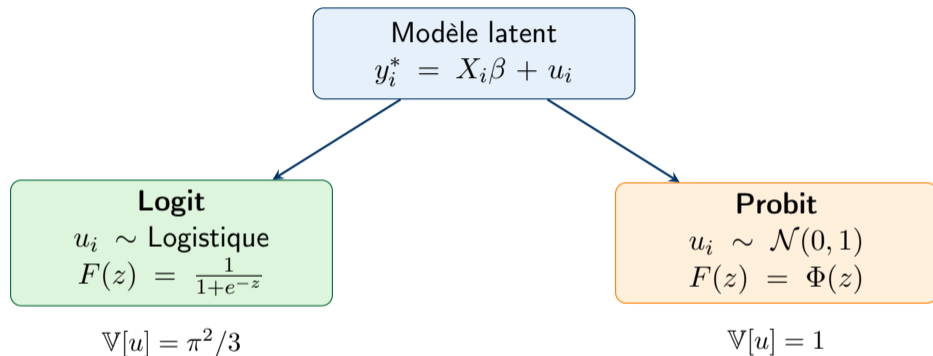
$$\begin{aligned} p_i &= \mathbb{P}(y_i = 1) = \mathbb{P}(y_i^* > 0) \\ &= \mathbb{P}(X_i\beta + u_i > 0) \\ &= \mathbb{P}(u_i > -X_i\beta) \\ &= 1 - F(-X_i\beta) \end{aligned}$$

où F est la fonction de répartition de u_i .

- Si $F(-z) = 1 - F(z)$ (loi symétrique), alors :

$$p_i = F(X_i\beta)$$

Cadre général – Choix de la distribution



Les deux lois sont symétriques. Le choix entre Logit et Probit est souvent empirique ; les résultats sont généralement similaires.

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Logit – Loi logistique

- ▶ La fonction de répartition de la loi logistique standard est :

$$F(z) = \frac{1}{1 + \exp(-z)} = \frac{\exp(z)}{1 + \exp(z)}$$

- ▶ Propriétés :

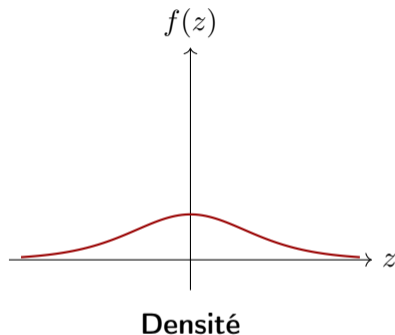
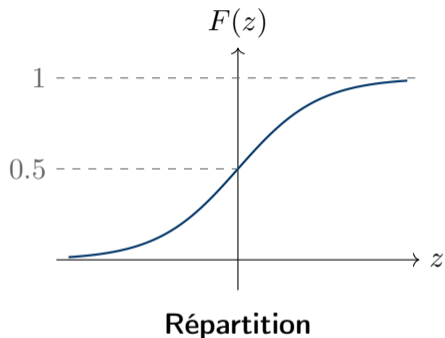
- ▶ **Symétrie** : $F(-z) = 1 - F(z)$

- ▶ **Densité** : $f(z) = F(z)(1 - F(z))$

- ▶ **Moments** : $\mathbb{E}[u] = 0$, $\mathbb{V}[u] = \frac{\pi^2}{3} \approx 3.29$

- ▶ La propriété $f(z) = F(z)(1 - F(z))$ simplifie considérablement les calculs de dérivées.

Logit – Graphique



La densité logistique ressemble à la densité normale mais avec des **queues plus épaisses** (plus de valeurs extrêmes).

Logit – Log-vraisemblance

- ▶ Avec $m_i = X_i\beta$:

$$p_i = F(m_i) = \frac{1}{1 + \exp(-m_i)}$$

- ▶ **Log-vraisemblance** :

$$\ell(y | X, \beta) = \sum_{i=1}^N [y_i \ln F(m_i) + (1 - y_i) \ln(1 - F(m_i))]$$

- ▶ La log-vraisemblance du Logit est **concave** \Rightarrow maximum unique, algorithmes standards convergent.

Logit – Score

- ▶ En utilisant $f(m_i) = F(m_i)(1 - F(m_i))$:

$$\begin{aligned} s(y | X, \beta) &= \sum_{i=1}^N X_i' \left[y_i \frac{f(m_i)}{F(m_i)} - (1 - y_i) \frac{f(m_i)}{1 - F(m_i)} \right] \\ &= \sum_{i=1}^N X_i' [y_i(1 - F(m_i)) - (1 - y_i)F(m_i)] \end{aligned}$$

- ▶ **Forme simplifiée :**

$$s(y | X, \beta) = \sum_{i=1}^N X_i' (y_i - F(m_i))$$

- ▶ Le score est une somme de résidus $(y_i - p_i)$ pondérés par X_i' .

- ▶ **Hessian :**

$$H(y | X, \beta) = \frac{\partial s}{\partial \beta'} = - \sum_{i=1}^N X_i' X_i f(m_i)$$

Comme $f(m_i) > 0$ pour tout i , on a $H \prec 0$ (défini négatif) \Rightarrow Newton-Raphson est **croissant**.

- ▶ **Information de Fisher :**

$$I(\beta) = \mathbb{E}[-H | X, \beta] = \sum_{i=1}^N X_i' X_i f(m_i)$$

- ▶ Le hessian ne dépend pas de $y \Rightarrow$ **Score = Newton-Raphson**.

- ▶ **BHHH :**

$$W_{BHHH}^{-1} = - \sum_{i=1}^N X_i' X_i (y_i - F(m_i))^2$$

- ▶ **Newton-Raphson (= Score) :**

$$W_{NR}^{-1} = - \sum_{i=1}^N X_i' X_i f(m_i)$$

- ▶ En pratique, Newton-Raphson converge très rapidement (3–5 itérations typiquement).

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Probit – Loi normale

- ▶ La fonction de répartition de la loi normale centrée réduite est :

$$F(z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds$$

- ▶ Propriétés :

- ▶ **Densité** : $\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$

- ▶ **Dérivée** : $\varphi'(z) = -z \cdot \varphi(z)$

- ▶ **Moments** : $\mathbb{E}[u] = 0, \mathbb{V}[u] = 1$

- ▶ La propriété $\varphi'(z) = -z\varphi(z)$ simplifie le calcul du hessien.

Probit – Log-vraisemblance

- ▶ Avec $m_i = X_i\beta$:

$$p_i = \Phi(m_i)$$

- ▶ **Log-vraisemblance** :

$$\ell(y | X, \beta) = \sum_{i=1}^N [y_i \ln \Phi(m_i) + (1 - y_i) \ln(1 - \Phi(m_i))]$$

- ▶ Comme pour le Logit, la log-vraisemblance est **concave**.

Probit – Score

- ▶ En notant $\varphi_i = \varphi(m_i)$ et $\Phi_i = \Phi(m_i)$:

$$s(y | X, \beta) = \sum_{i=1}^N X_i' \left[y_i \frac{\varphi_i}{\Phi_i} - (1 - y_i) \frac{\varphi_i}{1 - \Phi_i} \right]$$

- ▶ **Forme compacte :**

$$s(y | X, \beta) = \sum_{i=1}^N X_i' \frac{\varphi_i (y_i - \Phi_i)}{\Phi_i (1 - \Phi_i)}$$

- ▶ Contrairement au Logit, le score du Probit n'a pas de simplification aussi élégante.

Probit – Hessien et information de Fisher

► Hessien :

$$H = \sum_{i=1}^N X_i' X_i \varphi_i \left[\frac{(y_i - \Phi_i) m_i}{\Phi_i(1 - \Phi_i)} + \frac{(1 - 2\Phi_i) \varphi_i}{\Phi_i(1 - \Phi_i)} - \frac{\varphi_i}{\Phi_i(1 - \Phi_i)} \right]$$

► Information de Fisher :

$$I_1(\beta) = \mathbb{E}[-H \mid X, \beta] = \sum_{i=1}^N X_i' X_i \frac{\varphi_i^2}{\Phi_i(1 - \Phi_i)}$$

Différence avec le Logit

Le hessien **dépend de** $y \Rightarrow$ Score \neq Newton-Raphson.

Probit – Algorithmes

- ▶ **BHHH** :

$$W_{BHHH}^{-1} = - \sum_{i=1}^N X_i' X_i \frac{\varphi_i^2}{\Phi_i^2 (1 - \Phi_i)^2} (y_i - \Phi_i)^2$$

- ▶ **Score** :

$$W_{SC}^{-1} = - \sum_{i=1}^N X_i' X_i \frac{\varphi_i^2}{\Phi_i (1 - \Phi_i)}$$

- ▶ **Newton-Raphson** : expression complète du hessien.
- ▶ L'algorithme du Score est souvent préféré : moins de calculs que N-R, plus stable que BHHH.

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Comparaison – Identification

- ▶ Le vrai modèle latent s'écrit :

$$z_i^* = X_i b + v_i, \quad v_i \sim (0, \sigma^2)$$

- ▶ On estime en fait :

$$y_i^* = X_i \beta + u_i, \quad u_i \sim (0, 1)$$

avec $y_i^* = z_i^*/\sigma$, $\beta = b/\sigma$ et $u_i = v_i/\sigma$.

Non-identification

Les paramètres b et σ ne sont **pas identifiables séparément**. Seul $\beta = b/\sigma$ peut être estimé.

Comparaison – Interprétation des coefficients

► Ce qu'on peut faire :

1. Le **signe** de β_j est le même que celui de b_j ($\sigma > 0$)
2. Le **ratio** de deux coefficients : $\beta_j/\beta_k = b_j/b_k$
3. Comparer deux écarts tirés de la **même** régression

► Ce qu'on ne peut pas faire :

1. Comparer les coefficients de **deux régressions différentes**
2. Interpréter la **valeur absolue** d'un coefficient
3. Interpréter la **différence** entre deux coefficients

Comparaison – Conversion

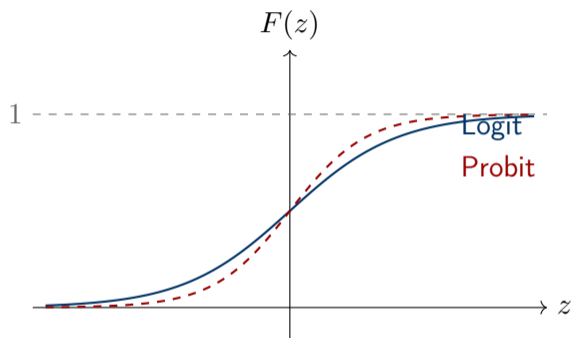
- ▶ **Probit** : $\mathbb{V}[u] = 1 \Rightarrow \beta_{\text{Probit}} = b/\sigma$.
- ▶ **Logit** : $\mathbb{V}[u] = \pi^2/3 \Rightarrow \beta_{\text{Logit}} = b/\phi$ avec $\phi = \sigma\sqrt{3}/\pi$.
- ▶ **Relation entre les coefficients** :

$$\beta_{\text{Probit}} \approx \frac{\sqrt{3}}{\pi} \beta_{\text{Logit}} \approx 0.55 \times \beta_{\text{Logit}}$$

$$\beta_{\text{Logit}} \approx 1.81 \times \beta_{\text{Probit}}$$

- ▶ Ces conversions sont **approximatives** car la loi logistique n'est pas exactement normale.

Comparaison – Graphique



Les deux courbes sont très proches. La logistique a des **queues plus épaisses** (coefficient d'aplatissement 1.2 vs 1 pour la normale).

Comparaison – Tableau

	Logit	Probit
Distribution	Logistique	Normale
Variance de u	$\pi^2/3 \approx 3.29$	1
Score = N-R ?	Oui	Non
Forme fermée de F	Oui	Non (intégrale)
Odds ratio simple	Oui	Non
Queues	Plus épaisses	Standard

En pratique, le choix importe peu pour les résultats. Le Logit est souvent préféré pour sa simplicité d'interprétation (odds ratio).

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Interprétation – Problème

- ▶ Les coefficients β ne sont définis qu'à une constante multiplicative près \Rightarrow **pas directement interprétables.**
- ▶ Trois solutions :
 1. **Odds ratio** : effet sur le rapport des probabilités
 2. **Effet marginal** : $\partial p / \partial X$
 3. **Effet incrémental** : variation de p entre deux valeurs de X
- ▶ On distingue les cas des variables explicatives **binaires** et **quantitatives**.

Interprétation – Odds

- ▶ Pour $X \in \{0, 1\}$, les **odds** (rapport des cotes) sont :

$$R(X) = \frac{\mathbb{P}(Y = 1 \mid X)}{\mathbb{P}(Y = 0 \mid X)} = \frac{F(\beta_0 + \beta_1 X)}{1 - F(\beta_0 + \beta_1 X)}$$

- ▶ L'**odds ratio** mesure l'effet de X :

$$\psi_X = \frac{R(1)}{R(0)} = \frac{\text{odds quand } X = 1}{\text{odds quand } X = 0}$$

- ▶ $\psi_X > 1$ signifie que $X = 1$ augmente les chances de $Y = 1$.

Interprétation – Odds ratio (Logit)

- ▶ Pour le modèle Logit :

$$R(X) = \frac{F(\beta_0 + \beta_1 X)}{1 - F(\beta_0 + \beta_1 X)} = \exp(\beta_0 + \beta_1 X)$$

- ▶ Odds ratio :

$$\psi_X = \frac{R(1)}{R(0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

- ▶ Il suffit de prendre l'**exponentielle du coefficient** pour obtenir l'odds ratio. Cette propriété est spécifique au Logit.

Interprétation – Effet incrémental

- ▶ L'**effet incrémental** de X sur la probabilité est la différence de probabilités quand X passe de 0 à 1 :

$$\Delta_X = \mathbb{P}(Y = 1 \mid X = 1) - \mathbb{P}(Y = 1 \mid X = 0) = F(\beta_0 + \beta_1) - F(\beta_0)$$

- ▶ Pour le Logit :

$$\Delta_X = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

- ▶ L'effet incrémental dépend du point de référence β_0 , contrairement à l'odds ratio.

Interprétation – Effet marginal

- ▶ L'effet marginal de X sur la probabilité est :

$$\delta_X = \frac{\partial \mathbb{P}(Y = 1 \mid X)}{\partial X} = \beta_1 \times f(\beta_0 + \beta_1 X)$$

- ▶ L'effet marginal **varie avec** X . On peut :
 - ▶ Évaluer au point moyen \bar{X}
 - ▶ Évaluer à la médiane
 - ▶ Tracer un graphique sur toutes les valeurs de X

Interprétation – Élasticité

- ▶ Pour une variable **continue** X_k , l'**élasticité** au point X mesure la variation relative de $\mathbb{P}(Y = 1)$ pour une variation de 1% de X_k :

$$e_{X_k} = \frac{\partial \ln \mathbb{P}(Y = 1 | X)}{\partial \ln X_k} = \beta_k \times f(X\beta) \times \frac{X_k}{F(X\beta)}$$

- ▶ L'élasticité permet de **comparer** l'influence de variables mesurées sur des échelles différentes (euros, années, heures, etc.).
- ▶ Comme l'effet marginal, l'élasticité varie avec X . On l'évalue généralement au **point moyen**.

Interprétation – Effet interdécile

- ▶ Mesurer la variation de probabilité quand X passe du 1er décile (D_1) au 9e décile (D_9) :

$$\Delta_{D_1 \rightarrow D_9} = F(\beta_0 + \beta_1 D_9) - F(\beta_0 + \beta_1 D_1)$$

- ▶ L'effet interdécile donne une bonne idée du **potentiel d'influence** de X sur la probabilité.
- ▶ Pour une loi normale, cela correspond à $X \pm 1.645\sigma_X$.

Interprétation – Odds ratio quantitatif (Logit)

- ▶ **Passage de a à b :**

$$\psi_X = \frac{R(b)}{R(a)} = \frac{\exp(\beta_0 + \beta_1 b)}{\exp(\beta_0 + \beta_1 a)} = \exp[\beta_1(b - a)]$$

- ▶ **Cas particulier :** augmentation d'une unité ($b = a + 1$) :

$$\boxed{\psi_X = \exp(\beta_1)}$$

- ▶ Pour le Logit, $\exp(\beta_1)$ est l'odds ratio associé à une augmentation d'une unité de X , que X soit binaire ou quantitative.

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Application 1 – Données

- ▶ **Source** : *Default of Credit Card Clients* (Yeh, 2009), UCI Machine Learning Repository.
- ▶ $N = 30\,000$ détenteurs de cartes de crédit d'une banque taïwanaise, observés entre avril et septembre 2005.
- ▶ **Variable expliquée** : défaut de paiement le mois suivant (octobre 2005). Codée 1 si défaut, 0 sinon.
- ▶ **Taux de défaut** : 22,1% (6 636 défauts sur 30 000).

Application 1 – Contexte

- ▶ Chaque client dispose d'une **limite de crédit** : montant maximal que la banque l'autorise à emprunter sur sa carte.
- ▶ C'est un attribut **structurel** du compte, fixé par la banque en fonction du profil du client (revenus, historique, etc.). Ce n'est pas un montant mensuel.
- ▶ Chaque mois, le client doit rembourser (au moins partiellement) son encours. Un **retard de paiement** est enregistré s'il ne respecte pas l'échéance.
- ▶ La variable LATE_SEPT vaut 1 si le client a au moins un mois de retard en septembre 2005. L'objectif est de prédire le **défaut en octobre**.

Terminologie

Le « défaut » désigne ici un **retard de paiement** (≥ 1 mois), pas une incapacité définitive de rembourser. En pratique bancaire, un vrai défaut (*charge-off*) n'est constaté qu'après plusieurs mois d'impayés (généralement 6 mois).

Application 1 – Variables explicatives

Variable	Type	Description
LIMIT_K	Quantitative	Limite de crédit accordée (en milliers de NT\$)
AGE	Quantitative	Âge du client (en années)
FEMALE	Binaire	= 1 si femme, = 0 si homme
MARRIED	Binaire	= 1 si marié(e), = 0 sinon
GRADUATE	Binaire	= 1 si diplôme de 3 ^e cycle
UNIVERSITY	Binaire	= 1 si diplôme universitaire (1 ^{er} /2 ^e cycle)
LATE_SEPT	Binaire	= 1 si retard de paiement \geq 1 mois en septembre

La catégorie de référence pour l'éducation est : lycée ou moins.

Application 1 – Statistiques descriptives

Variables quantitatives :

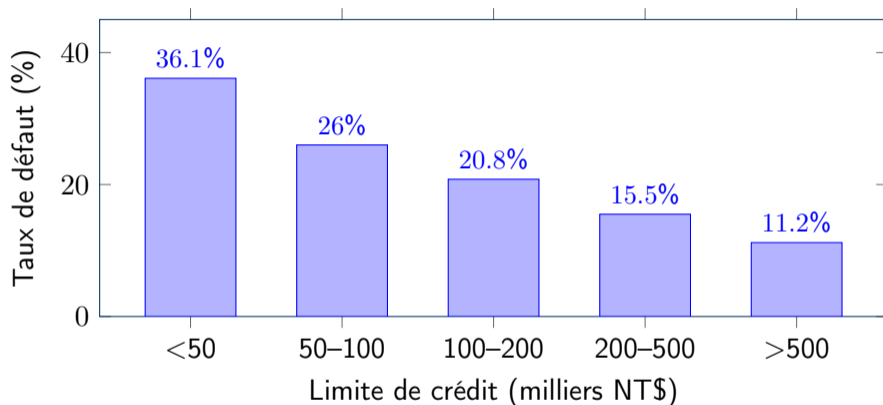
	LIMIT_K	AGE
Moyenne	167,5	35,5
Écart-type	129,7	9,2
Médiane	140	34
Min	10	21
Max	1 000	79

Variables binaires :

	Prop.	% déf.
Femme	60,4%	20,8%
Homme	39,6%	24,2%
Marié(e)	45,5%	23,5%
3 ^e cycle	35,3%	19,2%
Universitaire	46,8%	23,7%
Retard sept.	22,7%	50,3%

Le retard de paiement en septembre est le **signal le plus fort** : le taux de défaut passe de 13,8% (sans retard) à 50,3% (avec retard).

Application 1 – Taux de défaut par limite de crédit



Relation décroissante et monotone : plus la limite de crédit est élevée, plus le risque de défaut est faible.

Application 1 – Syntaxe Python

```
import statsmodels.api as sm

# Estimation Logit
logit = sm.Logit(y, X).fit()

# Estimation Probit
probit = sm.Probit(y, X).fit()

# Effets marginaux (variables continues)
mfx = logit.get_margeff(at="mean")

# Effets incrémentaux (variables binaires)
mfx_d = logit.get_margeff(at="mean", dummy=True)

# Élasticités (variables continues)
elas = logit.get_margeff(at="mean", method="eyex")
```

Application 1 – Résultats Logit et Probit

Variable	Logit		Probit	
	Coef.	(é.t.)	Coef.	(é.t.)
Constante	-1.668***	(0.081)	-0.999***	(0.046)
Limite crédit (milliers NT\$)	-0.0025***	(0.0001)	-0.0014***	(0.0001)
Femme	-0.158***	(0.031)	-0.088***	(0.018)
Marié(e)	+0.172***	(0.034)	+0.096***	(0.020)
Diplôme 3 ^e cycle	+0.080	(0.047)	+0.044	(0.027)
Diplôme universitaire	+0.120**	(0.042)	+0.065**	(0.024)
Âge	+0.005**	(0.002)	+0.003**	(0.001)
Retard paiement sept.	+1.751***	(0.031)	+1.042***	(0.019)
Log-vraisemblance	-13 822		-13 827	
<i>N</i>	30 000		30 000	

Le principal déterminant du défaut est le **retard de paiement**. Une limite de crédit élevée **réduit** le risque.

Application 1 – Conversion Logit → Probit

Variable	$\hat{\beta}_{\text{Logit}}$	$0.55 \times \hat{\beta}_{\text{Logit}}$	$\hat{\beta}_{\text{Probit}}$
Limite crédit	-0.0025	-0.0014	-0.0014
Femme	-0.158	-0.087	-0.088
Marié(e)	+0.172	+0.095	+0.096
Diplôme univ.	+0.120	+0.066	+0.065
Âge	+0.005	+0.003	+0.003
Retard sept.	+1.751	+0.965	+1.042

La conversion $\hat{\beta}_{\text{Probit}} \approx 0.55 \times \hat{\beta}_{\text{Logit}}$ fonctionne très bien, **sauf** pour le retard de paiement (coefficient élevé, queues de distribution).

Application 1 – Odds ratios (Logit)

Variable	$\hat{\beta}$	$\exp(\hat{\beta})$	Variation odds
Limite crédit (1000 NT\$)	-0.0025	0.998	-0.3%
Femme	-0.158	0.854	-14.6%
Marié(e)	+0.172	1.187	+18.7%
Diplôme 3 ^e cycle	+0.080	1.084	+8.4%
Diplôme universitaire	+0.120	1.128	+12.8%
Âge (1 an)	+0.005	1.005	+0.5%
Retard paiement sept.	+1.751	5.760	+476%

Un retard de paiement en septembre **multiplie par 5,8** les odds de défaut. Être femme les **réduit de 15%**.

Application 1 – Effets marginaux au point moyen

Variable	Logit	Probit
Limite crédit (1000 NT\$)	-0.0004***	-0.0004***
Âge (1 an)	+0.001**	+0.001**

L'effet marginal (dérivée) n'est défini que pour les variables **continues**. Pour les variables binaires, on utilise l'**effet incrémental**.

Application 1 – Élasticités au point moyen

Variable	Logit	Probit
Limite crédit	-0.341***	-0.325***
Âge	+0.146**	+0.145**

- ▶ Une hausse de 1% de la limite de crédit réduit la probabilité de défaut de **0,34%**.
- ▶ L'âge a une influence beaucoup plus faible : +0.15% pour une hausse de 1%.

Application 1 – Effets incrémentaux (Logit)

Variable	$\mathbb{P}(Y = 1 X = 1)$	$\mathbb{P}(Y = 1 X = 0)$	Δ
Femme	0.182	0.206	-0.025
Marié(e)	0.206	0.179	+0.027
Diplôme 3 ^e cycle	0.199	0.187	+0.013
Diplôme universitaire	0.201	0.183	+0.019
Retard paiement sept.	0.477	0.137	+0.341

Pour un individu « moyen », un retard de paiement fait passer la probabilité de défaut de **14%** à **48%**.

Application 1 – Effets interdéciles (Logit)

Variable	D_1	D_9	$\mathbb{P}(D_1)$	$\Delta_{D_1 \rightarrow D_9}$
Limite de crédit	30	360	0.250	-0.123
Âge	25	49	0.183	+0.019

- ▶ La limite de crédit a un **fort potentiel d'influence** : passer du 1^{er} au 9^e décile réduit la probabilité de défaut de **12,3 points**.
- ▶ L'âge a un effet **très faible** (+1.9 points entre 25 et 49 ans).

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Application 2 – Données

- ▶ **Source** : *Adult* (Becker & Kohavi, 1996), UCI Machine Learning Repository. Extrait du recensement américain de 1994.
- ▶ $N = 45\,222$ individus actifs (après suppression des valeurs manquantes).
- ▶ **Variable expliquée** : revenu annuel supérieur à 50 000 \$ (0/1).
- ▶ **Taux de revenu élevé** : 24,8%.

Application 2 – Contexte

- ▶ Le seuil de 50 000 \$ correspond environ au **75^e percentile** des revenus américains en 1994.
- ▶ On cherche à identifier les déterminants socio-démographiques d'un **revenu élevé** : éducation, âge, sexe, situation matrimoniale, temps de travail.
- ▶ Les variables d'éducation sont construites à partir d'une variable ordinale (`education_num`, de 1 à 16). On distingue trois niveaux :
 - ▶ **Bachelor+** : ≥ 13 (bachelor, master, doctorat)
 - ▶ **High school / College** : 9 à 12
 - ▶ **Référence** : ≤ 8 (sans diplôme du secondaire)

Application 2 – Variables explicatives

Variable	Type	Description
age	Quantitative	Âge (en années)
hours_per_week	Quantitative	Heures travaillées par semaine
FEMALE	Binaire	= 1 si femme, = 0 si homme
MARRIED	Binaire	= 1 si marié(e), = 0 sinon
WHITE	Binaire	= 1 si blanc, = 0 sinon
BACHELOR	Binaire	= 1 si diplôme \geq bachelor
HIGH_SCHOOL	Binaire	= 1 si high school ou college

La catégorie de référence pour l'éducation est : sans diplôme du secondaire.

Application 2 – Statistiques descriptives

Variables quantitatives :

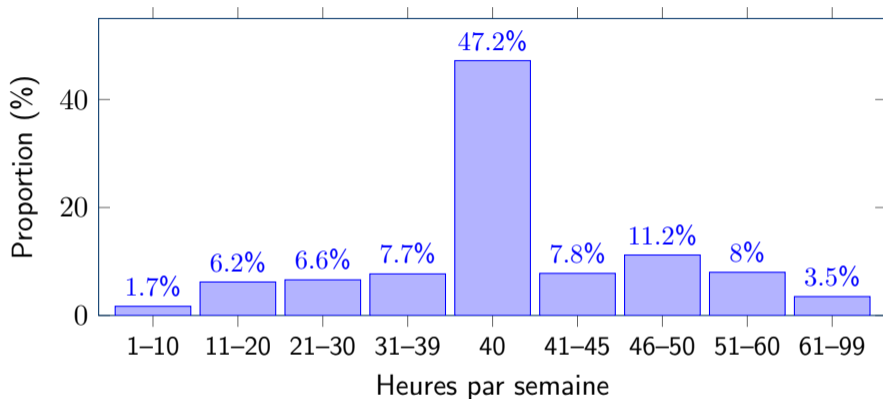
	age	hours
Moyenne	38,5	40,9
Écart-type	13,2	12,0
Médiane	37	40
Min	17	1
Max	90	99

Variables binaires :

	Prop.	% >50K
Femme	32,5%	11,4%
Homme	67,5%	31,2%
Marié(e)	46,6%	45,4%
Blanc	86,0%	26,2%
Bachelor+	25,2%	48,7%
High school	62,2%	18,9%

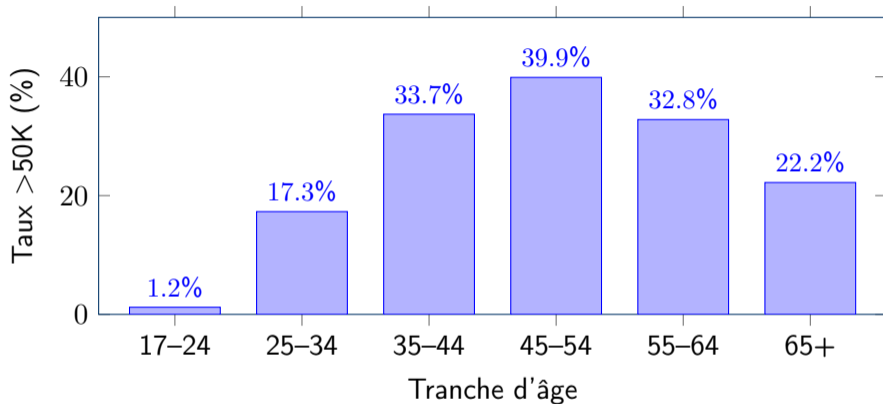
L'éducation et le statut matrimonial sont les variables les plus discriminantes : 48,7% des diplômés Bachelor+ gagnent >50K contre 18,9% pour les High school.

Application 2 – Distribution des heures travaillées



47% des individus déclarent exactement 40 heures (semaine légale standard aux É.-U.). La variable n'est pas véritablement continue : l'effet marginal et l'élasticité ont ici un intérêt limité. L'effet **interdécile** (25h → 55h) est plus informatif.

Application 2 – Taux de revenu élevé par âge



Relation en **cloche** : le taux de revenu élevé croît avec l'âge jusqu'à 45-54 ans, puis décroît. Cela justifie un modèle linéaire en âge comme approximation sur la population active.

Application 2 – Résultats Logit et Probit

Variable	Logit		Probit	
	Coef.	(é.t.)	Coef.	(é.t.)
Constante	-7.260***	(0.111)	-4.119***	(0.060)
Âge	+0.032***	(0.001)	+0.019***	(0.001)
Femme	-0.085*	(0.038)	-0.074***	(0.021)
Marié(e)	+2.288***	(0.035)	+1.283***	(0.019)
Blanc	+0.246***	(0.044)	+0.133***	(0.025)
Diplôme \geq Bachelor	+3.016***	(0.065)	+1.694***	(0.034)
High school / College	+1.497***	(0.062)	+0.819***	(0.033)
Heures / semaine	+0.032***	(0.001)	+0.018***	(0.001)
Log-vraisemblance	-17 358		-17 314	
<i>N</i>	45 222		45 222	

Tous les coefficients sont significatifs. Les principaux déterminants du revenu élevé sont le **diplôme** et le **statut matrimonial**.

Application 2 – Conversion Logit → Probit

Variable	$\hat{\beta}_{\text{Logit}}$	$0.55 \times \hat{\beta}_{\text{Logit}}$	$\hat{\beta}_{\text{Probit}}$
Âge	+0.032	+0.018	+0.019
Femme	-0.085	-0.047	-0.074
Marié(e)	+2.288	+1.262	+1.283
Blanc	+0.246	+0.136	+0.133
Diplôme \geq Bachelor	+3.016	+1.663	+1.694
High school / College	+1.497	+0.826	+0.819
Heures / semaine	+0.032	+0.017	+0.018

La conversion $\times 0.55$ fonctionne bien pour la plupart des variables. L'écart pour FEMALE (coefficient faible, peu significatif en Logit) est plus marqué.

Application 2 – Odds ratios (Logit)

Variable	$\hat{\beta}$	$\exp(\hat{\beta})$	Variation odds
Âge (1 an)	+0.032	1.033	+3.3%
Femme	-0.085	0.919	-8.1%
Marié(e)	+2.288	9.857	+886%
Blanc	+0.246	1.279	+27.9%
Diplôme \geq Bachelor	+3.016	20.407	+1 941%
High school / College	+1.497	4.470	+347%
Heures / semaine (1h)	+0.032	1.032	+3.2%

Un diplôme Bachelor+ **multiplie par 20** les odds de revenu élevé. Le mariage les **multiplie par 10** (effet de sélection probable).

Application 2 – Effets marginaux au point moyen

Variable	Logit	Probit
Âge (1 an)	+0.004***	+0.004***
Heures / semaine (1h)	+0.004***	+0.004***

L'effet marginal (dérivée) n'est défini que pour les variables **continues**. Pour les variables binaires, on utilise l'**effet incrémental**.

Application 2 – Élasticités au point moyen

Variable	Logit	Probit
Âge	+1.082***	+1.134***
Heures / semaine	+1.106***	+1.156***

- ▶ Les deux variables continues ont des élasticités proches de 1 : une hausse de 1% de l'âge ou des heures travaillées augmente la probabilité de revenu élevé d'environ 1%.
- ▶ Les élasticités permettent de comparer des variables sur des échelles différentes : ici l'âge et les heures ont un effet **comparable**.

Application 2 – Effets incrémentaux (Logit)

Variable	$\mathbb{P}(Y = 1 X = 1)$	$\mathbb{P}(Y = 1 X = 0)$	Δ
Femme	0.136	0.146	-0.010
Marié(e)	0.361	0.054	+0.307
Blanc	0.147	0.119	+0.028
Diplôme \geq Bachelor	0.614	0.072	+0.541
High school / College	0.227	0.062	+0.165

Pour un individu « moyen », un diplôme Bachelor+ fait passer la probabilité de revenu élevé de **7% à 61%**.

Application 2 – Effets interdéciles (Logit)

Variable	D_1	D_9	$\mathbb{P}(D_1)$	$\Delta_{D_1 \rightarrow D_9}$
Âge	22	57	0.089	+0.142
Heures / semaine	25	55	0.092	+0.114

- ▶ L'âge a un **potentiel d'influence notable** : +14.2 points entre 22 et 57 ans.
- ▶ Les heures travaillées ont un effet comparable : +11.4 points entre 25 et 55 heures/semaine.

Plan

Cadre général

Le modèle Logit

Le modèle Probit

Comparaison Logit / Probit

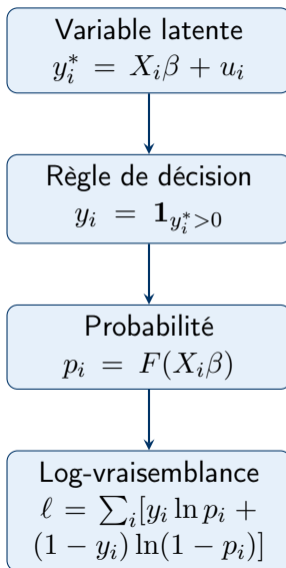
Aides à l'interprétation

Application 1

Application 2

Récapitulatif

Récapitulatif – Modèle général



Récapitulatif – Formules

► **Logit :**

$$F(z) = \frac{1}{1 + e^{-z}}, \quad f(z) = F(z)(1 - F(z))$$
$$s = \sum_i X_i'(y_i - F_i), \quad H = - \sum_i X_i'X_i f_i$$

► **Probit :**

$$F(z) = \Phi(z), \quad f(z) = \varphi(z)$$
$$s = \sum_i X_i' \frac{\varphi_i(y_i - \Phi_i)}{\Phi_i(1 - \Phi_i)}, \quad I = \sum_i X_i'X_i \frac{\varphi_i^2}{\Phi_i(1 - \Phi_i)}$$

► **Conversion :** $\beta_{\text{Probit}} \approx 0.55 \times \beta_{\text{Logit}}$.

Récapitulatif – Points clés

1. Les modèles dichotomiques reposent sur une **variable latente**.
2. Les coefficients ne sont identifiés qu'à σ près.
3. **Logit** : odds ratio simple $\exp(\beta)$, Score = N-R.
4. **Probit** : variance normalisée à 1, trois algorithmes.
5. Pour interpréter : **odds ratio, effet marginal, effet incrémental**.
6. En pratique, Logit et Probit donnent des résultats **très similaires**.

Récapitulatif – Références

- ▶ Greene, W. (2018). *Econometric Analysis*, 8th ed., Pearson, ch. 17.
- ▶ Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, ch. 15.
- ▶ Train, K. (2009). *Discrete Choice Methods with Simulation*, Cambridge.
- ▶ Cameron, A.C. & Trivedi, P.K. (2005). *Microeconometrics: Methods and Applications*, Cambridge University Press, ch. 16.