Économétrie des Variables Qualitatives

Introduction

Janvier, 2026

Plan du cours

— Variables explicatives

Variables qualitatives **à droite** de l'équation

- Modèle sans terme constant
- Modèle avec terme constant
- Modèle avec variables explicatives
- Modèle avec produits croisés

PARTIE II — Variables expliquées

Variables qualitatives **à gauche** de l'équation

- Variables dichotomiques
- Variables polytomiques ordonnées
- Variables de comptage
- Variables censurées/tronquées

$$y_i = f(D_i, X_i) + u_i$$

Partie I

$$y_i \in \{0, 1, \ldots\}$$

Partie II

PARTIE I

Les variables qualitatives EXPLICATIVES

Variables qualitatives à droite de l'équation

$$y_i = f(D_i, X_i) + u_i$$

Introduction

Contexte

Les variables qualitatives explicatives sont omniprésentes en économie appliquée :

- Économie du travail : genre, niveau de diplôme, catégorie socio-professionnelle
- Économie de l'innovation : secteur d'activité, région, taille d'entreprise
- Économie industrielle : appartenance à un groupe, type de marché

Objectif

Comprendre l'interprétation des coefficients des variables qualitatives dans le modèle linéaire et ses extensions.

Deux utilisations principales

1. **Effets fixes catégoriels** : indicatrices d'appartenance à un groupe
   - Les coefficients s'interprètent comme des écarts moyens par rapport à une modalité de référence
   - Ils ne représentent plus des dérivées (qui n'existent pas)

2. **Approximation de fonctions non linéaires** :
   - Découpage d'une variable continue en intervalles
   - Estimation d'une relation non paramétrique par morceaux

Modèle sans terme constant : Formalisation

Définition (Variable qualitative polytomique)

Soit une variable qualitative à p modalités. Pour un échantillon de N individus, on définit les ensembles d'indices :

$$G_j = \{i : \text{individu } i \text{ appartient au groupe } j\}, \quad j = 1, \dots, p$$

avec $\bigcup_{j=1}^{p} G_j = \{1, \dots, N\}$ (partition).

Définition (Indicatrices)

Les variables dichotomiques associées sont définies par :

$$D_{ji} = \begin{cases} 1 & \text{si } i \in G_j \\ 0 & \text{si } i \notin G_j \end{cases}, \quad i = 1, \dots, N$$

Propriétés fondamentales des indicatrices

Propriété

Les variables indicatrices vérifient les propriétés suivantes :

1. **Idempotence** : $D_{ji}^2 = D_{ji} \quad (\text{car } 0^2 = 0 \text{ et } 1^2 = 1)$
2. **Exclusivité mutuelle** : $D_{ji} \cdot D_{ki} = 0$ pour tout $j \neq k$
3. **Effectif du groupe** : $\sum_{i=1}^{N} D_{ji} = N_j$
4. **Fréquence du groupe** : $\frac{1}{N} \sum_{i=1}^{N} D_{ji} = \frac{N_j}{N}$

Remarque

La propriété 4 montre que pour les indicatrices, la moyenne arithmétique calcule des pourcentages.

Le modèle linéaire avec indicatrices

Spécification

Le modèle linéaire s'écrit :

$$y_i = \sum_{j=1}^{p} b_j D_{ji} + u_i, \quad i = 1, \ldots, N$$

avec les hypothèses classiques sur les perturbations :

$$\mathbb{E}(u_i) = 0, \quad \mathbb{E}(u_i^2) = \sigma_u^2, \quad \mathbb{E}(u_i u_j) = 0 \text{ pour } i \neq j$$

Proposition (Interprétation des coefficients)

L'espérance conditionnelle dans le groupe j est :

$$\mathbb{E}(y_i \mid D) = b_j \quad \text{si } i \in G_j$$

Les coefficients représentent donc les moyennes conditionnelles par groupe.

Différence entre groupes

Corollaire

La différence entre deux coefficients s'interprète comme la différence des espérances conditionnelles :

$$b_j - b_k = \mathbb{E}(y_i \mid i \in G_j) - \mathbb{E}(y_i \mid i \in G_k)$$

Distinction avec les variables quantitatives

- Variables **quantitatives** : $b_j = \frac{\partial \mathbb{E}(y)}{\partial x_j}$ (dérivée)
- Variables **qualitatives** : $b_j = \mathbb{E}(y \mid \text{groupe } j)$ (moyenne conditionnelle)

Écriture matricielle

Notation

Pour chaque individu i, on définit le vecteur ligne :

$$D_i = (D_{1i}, D_{2i}, \ldots, D_{pi})_{1\times p}$$

et le vecteur des paramètres :

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}_{p\times 1}$$

Le modèle individuel s'écrit :

$$y_i = D_i b + u_i, \quad i = 1, \ldots, N$$

Estimateur des MCO : Dérivation (1/2)

Formule générale

L'estimateur des MCO est :

$$\hat{b} = \left(\sum_{i=1}^{N} D_i' D_i \right)^{-1} \sum_{i=1}^{N} D_i' y_i$$

Calcul de $D_i' D_i$

En utilisant l'idempotence et l'exclusivité mutuelle :

$$D_i' D_i = \begin{pmatrix} D_{1i} \\ \vdots \\ D_{pi} \end{pmatrix} (D_{1i}, \ldots, D_{pi}) = \begin{pmatrix} D_{1i} & 0 & \cdots & 0 \\ 0 & D_{2i} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_{pi} \end{pmatrix}$$

Estimateur des MCO : Dérivation (2/2)

Somme sur les individus

$$\sum_{i=1}^{N} D_i' D_i = \begin{pmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & N_p \end{pmatrix}$$

où $N_j = \sum_{i=1}^{N} D_{ji}$ est l'effectif du groupe j.

Théorème (Estimateur MCO)

L'estimateur des MCO des coefficients est :

$$\hat{b}_j = \frac{1}{N_j} \sum_{i \in G_j} y_i = \bar{y}_j$$

C'est-à-dire la moyenne arithmétique de y dans chaque groupe.

Démonstration détaillée

Démonstration.

Pour le second membre de l'estimateur :

$$\sum_{i=1}^{N} D_i' y_i = \begin{pmatrix} \sum_{i=1}^{N} D_{1i} y_i \\ \vdots \\ \sum_{i=1}^{N} D_{ji} y_i \\ \vdots \\ \sum_{i=1}^{N} D_{pi} y_i \end{pmatrix} = \begin{pmatrix} \sum_{i \in G_1} y_i \\ \vdots \\ \sum_{i \in G_j} y_i \\ \vdots \\ \sum_{i \in G_p} y_i \end{pmatrix}$$

En combinant avec l'inverse de la matrice diagonale :

$$\hat{b} = \begin{pmatrix} 1/N_1 & & 0 \\ & \ddots & \\ 0 & & 1/N_p \end{pmatrix} \begin{pmatrix} \sum_{i \in G_1} y_i \\ \vdots \\ \sum_{i \in G_p} y_i \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_p \end{pmatrix}$$

Le problème de la multicolinéarité parfaite

Observation clé

Le terme constant e (vecteur unitaire) est égal à la somme des indicatrices :

$$e = \sum_{j=1}^{p} D_j$$

car tout individu appartient exactement à un groupe.

Conséquence

Dans un modèle avec terme constant, il faut retirer une indicatrice pour éviter la multicolinéarité parfaite. Cette modalité devient la **modalité de référence**.

Modèle avec modalité de référence

Spécification

Si l'on retire la modalité k, le modèle devient :

$$y_i = c_0 + \sum_{j \neq k} c_j D_{ji} + u_i$$

Proposition (Relation avec le modèle complet)

Les coefficients c sont reliés aux coefficients b du modèle sans constante par :

$$\begin{aligned} c_0 &= b_k \quad \text{(coefficient de la modalité de référence)} \\ c_j &= b_j - b_k \quad \text{pour } j \neq k \quad \text{(écart à la référence)} \end{aligned}$$

Démonstration

Démonstration.

La prévision du modèle avec constante s'écrit :

$$\hat{y} = c_0 e + \sum_{j \neq k} c_j D_j$$

En remplaçant $e = \sum_{j=1}^{p} D_j$:

$$\begin{aligned}\hat{y} &= c_0 \sum_{j=1}^{p} D_j + \sum_{j \neq k} c_j D_j \\ &= (c_0 + c_1) D_1 + \cdots + c_0 D_k + \cdots + (c_0 + c_p) D_p\end{aligned}$$

La prévision du modèle sans constante est $\hat{y} = \sum_{j=1}^{p} b_j D_j$.
Par unicité de la décomposition, on identifie :

$$c_0 + c_j = b_j \text{ pour } j \neq k \quad \text{et} \quad c_0 = b_k$$

Interprétation des coefficients

Résumé

Coefficient	Interprétation
c_0	Moyenne du groupe de référence k : $\mathbb{E}(y_i \mid i \in G_k)$
c_j	Écart par rapport à la référence : $\mathbb{E}(y_i \mid i \in G_j) - \mathbb{E}(y_i \mid i \in G_k)$

Importance pratique

Il est indispensable d'indiquer explicitement la modalité de référence retirée dans les tableaux de régression pour permettre une interprétation correcte des résultats.

Remarque (Test de Fisher)

Le test de Fisher sur le modèle avec terme constant teste l'hypothèse :

$$H_0 : c_1 = c_2 = \cdots = c_{k-1} = c_{k+1} = \cdots = c_p = 0$$

Ce qui équivaut à :

$$H_0 : \mathbb{E}(y_i \mid i \in G_j) = \mathbb{E}(y_i \mid i \in G_k) \quad \forall j \neq k$$

C'est un test d'égalité des moyennes entre tous les groupes.

Remarque (Test de Student)

Un test de Student sur c_j teste l'égalité des moyennes entre le groupe j et le groupe de référence k.

Extension avec variables quantitatives

Spécification

On introduit une matrice de variables explicatives X_i :

$$y_i = X_i a + D_i b + u_i$$

où X_i est de dimension $(1 \times m)$ et D_i de dimension $(1 \times p)$.

Proposition (Espérance conditionnelle)

L'espérance conditionnelle dans le groupe j est :

$$\mathbb{E}(y_i \mid X_i, D_{ji} = 1) = X_i a + b_j$$

Interprétation

Corollaire

La différence entre les groupes j et k, à X fixé, est :

$$\begin{aligned} &\mathbb{E}(y_i \mid X_i, D_{ji} = 1) - \mathbb{E}(y_i \mid X_i, D_{ki} = 1) \\ &= (X_i a + b_j) - (X_i a + b_k) = b_j - b_k \end{aligned}$$

Conclusion

Les résultats précédents restent valables :

- Le terme constant représente le coefficient de l'indicatrice retirée
- Les autres coefficients mesurent l'écart à cette référence
- Ces écarts sont contrôlés pour les autres variables X

Motivation : Évaluation d'une politique

Contexte

On étudie l'effet d'une mesure d'aide (affectée au hasard) sur une variable de performance y_i.

Définition (Indicatrice de traitement)

$$D_i = \begin{cases} 1 & \text{si l'individu } i \text{ est aidé} \\ 0 & \text{sinon} \end{cases}$$

Résultats potentiels

Pour chaque individu, deux résultats potentiels :

- y_{0i} : performance si l'individu i n'est pas aidé
- y_{1i} : performance si l'individu i est aidé

Effet causal du traitement

Définition (Effet moyen du traitement)

L'effet causal recherché est :

$$\delta = \mathbb{E}(y_{1i} - y_{0i})$$

C'est la moyenne des variations de performance associées à la mesure.

Modélisation des résultats potentiels

On suppose :

$$y_{0i} = a_0 + X_i c_0 + u_{0i}$$
$$y_{1i} = a_1 + X_i c_1 + u_{1i}$$

où X_i représente les déterminants de la performance.

Variable observable

Problème fondamental de l'inférence causale

On n'observe que l'un des deux résultats potentiels :

$$y_i = D_i \cdot y_{1i} + (1 - D_i) \cdot y_{0i} = \begin{cases} y_{1i} & \text{si } D_i = 1 \\ y_{0i} & \text{si } D_i = 0 \end{cases}$$

Modèle économétrique

En substituant :

$$\begin{aligned} y_i &= D_i(a_1 + X_i c_1 + u_{1i}) + (1 - D_i)(a_0 + X_i c_0 + u_{0i}) \\ &= a_0 + X_i c_0 + D_i \underbrace{(a_1 - a_0)}_{a} + D_i X_i \underbrace{(c_1 - c_0)}_{c} + u_i \end{aligned}$$

Modèle avec interactions

Théorème (Modèle avec produits croisés)

Le modèle économétrique s'écrit :

$$y_i = a_0 + X_i c_0 + D_i \cdot a + (D_i \otimes X_i) \cdot c + u_i$$

où :

- $a = a_1 - a_0$: effet du traitement à $X = 0$
- $c = c_1 - c_0$: modification des pentes par le traitement
- $u_i = D_i u_{1i} + (1 - D_i) u_{0i}$: perturbation composite

Hétérogénéité des effets

Ce modèle autorise une hétérogénéité de l'effet du traitement selon les caractéristiques X_i.

Rappel : Le produit de Kronecker

Définition (Produit de Kronecker)

Soient $A = (a_{ij})$ une matrice de dimension $(m \times n)$ et B une matrice de dimension $(p \times q)$. Le produit de Kronecker $A \otimes B$ est la matrice de dimension $(mp \times nq)$ définie par blocs :

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

Exemple (Application aux indicatrices)

Avec $D_i = (D_{1i}, \ldots, D_{pi})$ vecteur $(1 \times p)$ et X_i vecteur $(1 \times m)$:

$$D_i \otimes X_i = (D_{1i}X_i, D_{2i}X_i, \ldots, D_{pi}X_i)_{1\times pm}$$

Chaque bloc $D_{ji}X_i$ correspond aux variables X multipliées par l'indicatrice du groupe j.

Calcul de l'effet moyen

Proposition

L'effet du traitement est :

$$\delta = \mathbb{E}(y_{1i} - y_{0i}) = (a_1 - a_0) + \mathbb{E}(X)(c_1 - c_0) = a + \mathbb{E}(X)c$$

Estimation

Un estimateur convergent est :

$$\hat{\delta} = \hat{a} + \bar{X}\hat{c}$$

Astuce pratique

Si les variables X sont centrées avant de calculer les produits croisés (i.e., $\bar{X} = 0$), alors :

$$\hat{\delta} = \hat{a}$$

L'effet moyen est directement donné par le coefficient de l'indicatrice.

Hétéroscédasticité induite

Remarque (Structure de la variance)

La perturbation composite $u_i = D_i u_{1i} + (1 - D_i) u_{0i}$ implique :

$$\mathbb{V}[u_i] = \begin{cases} \mathbb{V}[u_{0i}] & \text{si } D_i = 0 \\ \mathbb{V}[u_{1i}] & \text{si } D_i = 1 \end{cases}$$

Conséquence

Si $\mathbb{V}[u_{0i}] \neq \mathbb{V}[u_{1i}]$, le modèle présente une hétéroscédasticité par bloc. Dans ce cas :

- Les MCO restent convergents mais inefficaces
- Il faut utiliser les **moindres carrés pondérés** ou les **écarts-types robustes**

Cas polytomique

Extension

Avec p groupes et des produits croisés :

$$y_i = D_i b + (X_i \otimes D_i) c + u_i$$

où le terme en X seul est retiré car $\sum_{j=1}^{p} X_i D_{ji} = X_i$.

Proposition

L'espérance conditionnelle dans le groupe j devient :

$$\mathbb{E}(y_i \mid X_i, D_{ji} = 1) = X_i c_j + b_j$$

La différence entre groupes j et k est :

$$\gamma_i = X_i (c_j - c_k) + (b_j - b_k)$$

L'effet varie selon les caractéristiques individuelles X_i.

Effet moyen et centrage

Théorème

L'effet moyen est :

$$\bar{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i = \bar{X}(c_j - c_k) + (b_j - b_k)$$

Corollaire (Méthode du centrage)

Si l'on centre X avant de calculer les produits croisés ($\bar{X} = 0$), alors :

$$\bar{\gamma} = b_j - b_k$$

La différence de coefficients mesure directement l'écart moyen entre groupes, une fois éliminé l'effet des variables de X.

PARTIE II

Les variables qualitatives EXPLIQUÉES

Variables qualitatives à gauche de l'équation

$$y_i \in \{0, 1, \ldots\} \quad \longleftarrow \quad f(X_i, \theta)$$

Introduction

Contexte

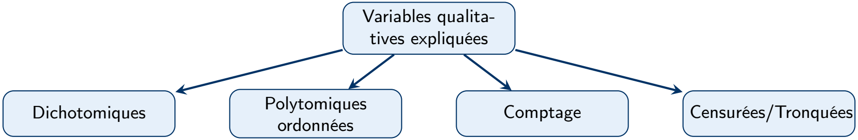
Les bases de données microéconomiques contiennent fréquemment :

- Des données tronquées ou censurées
- Des informations connues seulement par intervalle
- Des variables purement qualitatives

Exemple

- Enquête Innovation (SESSI) : fait d'avoir innové, importance d'un déterminant
- Enquête Emploi (INSEE) : statut d'emploi, heures travaillées (observées seulement si emploi)

Classification des variables qualitatives expliquées



Définition et codage

Définition (Variable dichotomique)

Une variable dichotomique ne peut prendre que deux modalités exclusives l'une de l'autre (Oui/Non, Supérieur/Inférieur à un seuil, etc.).

Par convention, on code :

$$y_i = \begin{cases} 1 & \text{si l'événement se produit} \\ 0 & \text{sinon} \end{cases}$$

Exemple (Innovation)

$$y_i = \begin{cases} 1 & \text{si l'entreprise } i \text{ a innové} \\ 0 & \text{sinon} \end{cases}$$

Objectif économétrique

Ce que l'on cherche à expliquer

Les déterminants de la probabilité que l'événement se produise :

$$\mathbb{P}(y_i = 1 \mid X_i) = ?$$

On cherche les variables qui augmentent ou réduisent cette probabilité.

Idée

On va construire un **modèle latent** (inobservable) qui représente le critère de décision sous-jacent.

Le cadre : variable dépendante binaire

Objectif : Modéliser la probabilité qu'un événement se réalise.

Exemples :

- Participer au marché du travail ($Y_i = 1$) ou non ($Y_i = 0$)
- Acheter un produit, faire défaut sur un crédit, voter pour un candidat...

Modèle de Probabilité Linéaire (MPL) :

$$Y_i = X_i'\beta + \varepsilon_i \quad \text{avec} \quad Y_i \in \{0, 1\}$$

On a :

$$E[Y_i|X_i] = X_i'\beta = P(Y_i = 1|X_i)(\text{loi de Bernoulli})$$

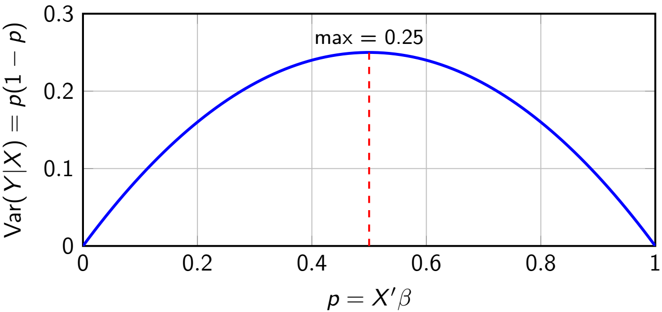
Question : Peut-on estimer β par MCO comme dans le modèle linéaire standard ?

Problème 1 : Hétéroscédasticité structurelle

$Y_i|X_i$ suit une loi de **Bernoulli** de paramètre $p_i = X_i'\beta$.

Variance conditionnelle :

$$\mathrm{Var}(Y_i|X_i) = p_i(1 - p_i) = X_i'\beta \cdot (1 - X_i'\beta)$$



$\Rightarrow$ La variance **dépend de** X_i : hétéroscédasticité **inhérente** au modèle.

Conséquences de l'hétéroscédasticité

Rappel : Sous hétéroscédasticité, l'estimateur MCO reste sans biais et convergent, mais :

1. **Inefficacité :** $\hat{\beta}_{MCO}$ n'est plus BLUE (Best Linear Unbiased Estimator)

2. **Inférence invalide :** La matrice de variance usuelle

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$$

   est **incorrecte**. Les tests t et F ne sont plus valides.

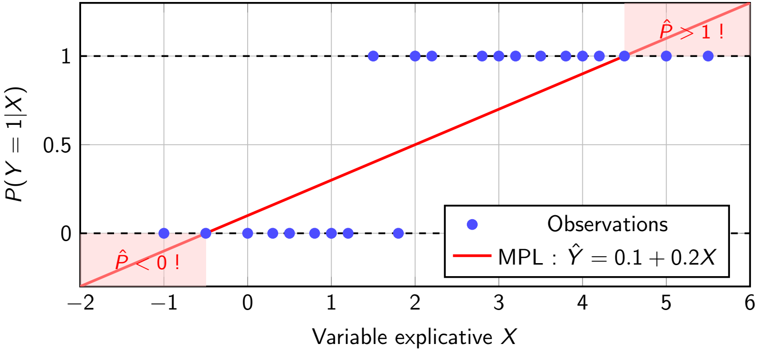
3. **Solution partielle :** Utiliser les écarts-types robustes de White (HC)

$$\widehat{\text{Var}}_{HC}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i'\right)(X'X)^{-1}$$

Mais cela ne résout pas les autres problèmes...

Problème 2 : Prédictions hors de $[0, 1]$

Le modèle linéaire n'impose **aucune contrainte** sur $\hat{Y}_i = X_i'\hat{\beta}$.



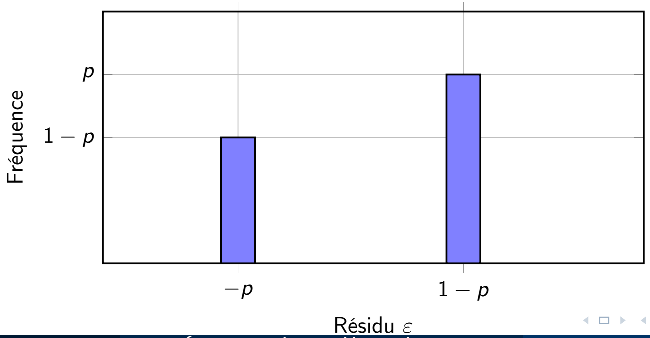
$\Rightarrow$ Des **probabilités négatives** ou **supérieures à 1** : absurde !

Problème 3 : Non-normalité des résidus

Les résidus ne peuvent prendre que **deux valeurs** :

$$\varepsilon_i = Y_i - X_i'\beta = \begin{cases} 1 - X_i'\beta & \text{si } Y_i = 1 \\ -X_i'\beta & \text{si } Y_i = 0 \end{cases}$$

Distribution des résidus (exemple : $p = X'\beta = 0.6$)



Problème 4 : Effets marginaux constants (irréalistes)

Dans le MPL, l'effet marginal est **constant** :

$$\frac{\partial P(Y=1|X)}{\partial X_k} = \beta_k \quad \forall X$$

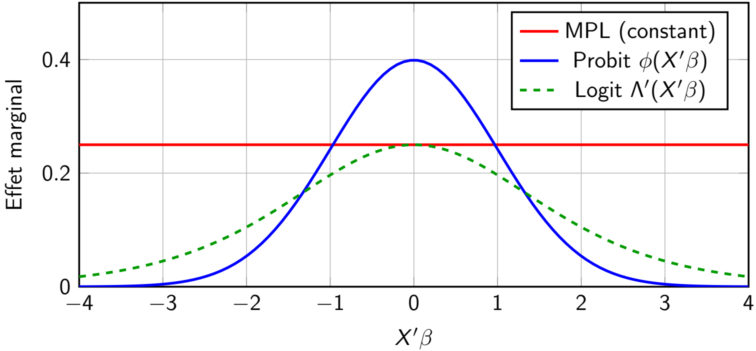
Problème conceptuel :

- Passer de $P = 0.01$ à $P = 0.06$: effet faible (événement rare reste rare)
- Passer de $P = 0.47$ à $P = 0.52$: effet fort (basculement de majorité)
- Passer de $P = 0.94$ à $P = 0.99$: effet faible (quasi-certitude confirmée)

$\Rightarrow$ L'effet d'une variable devrait **s'atténuer aux extrêmes** (saturation).

Intuition : Augmenter le revenu de quelqu'un qui a déjà 99% de chances d'acheter une maison n'aura presque aucun effet sur sa probabilité d'achat.

Comparaison des effets marginaux



Modèles non-linéaires : L'effet marginal est maximal au centre ($P \approx 0.5$) et tend vers 0 aux extrêmes $\Rightarrow$ **effet de saturation** réaliste.

Récapitulatif : limites du MPL

Problème	**Conséquence**	**Gravité**
Hétéroscédasticité structurelle	Inférence invalide (corrigeable par HC)	Moyenne
Prédictions hors $[0, 1]$	Probabilités absurdes	Élevée
Résidus non-normaux	Tests invalides (asymptotiquement OK)	Moyenne
Effets marginaux constants	Modèle mal spécifié	Élevée

Conclusion : Le MPL peut être utilisé comme **approximation rapide** pour des probabilités proches de 0.5, mais il est **inadapté** pour une modélisation rigoureuse.

Solution : les modèles à variable latente

Idée : Introduire une variable latente Y_i^* continue :

$$Y_i^* = X_i'\beta + \varepsilon_i, \qquad Y_i = 1(Y_i^* > 0)$$

Si $\varepsilon_i \sim F$ (fonction de répartition), alors :

$$P(Y_i = 1|X_i) = P(Y_i^* > 0|X_i) = P(\varepsilon_i > -X_i'\beta) = F(X_i'\beta)$$

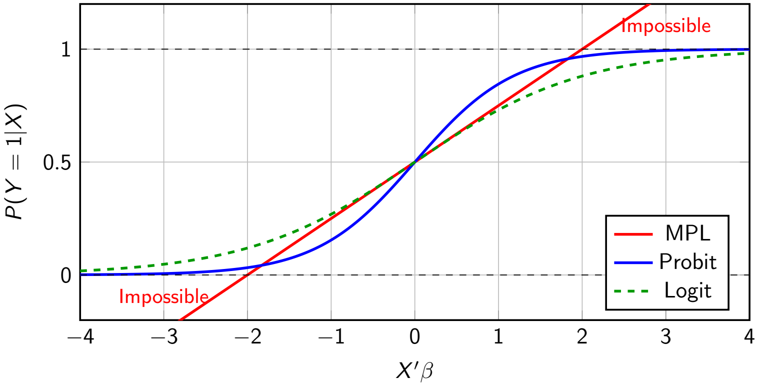
Choix classiques :

- $F = \Phi$ (normale standard) $\Rightarrow$ **Modèle Probit**
- $F = \Lambda$ (logistique) $\Rightarrow$ **Modèle Logit**

Propriétés :

1. $F(X'\beta) \in [0, 1]$ par construction
2. Effets marginaux décroissants aux extrêmes : $\frac{\partial P}{\partial X_k} = f(X'\beta) \cdot \beta_k$
3. Estimation par **maximum de vraisemblance**

Comparaison graphique : MPL vs Probit/Logit



$\Rightarrow$ Les modèles Probit et Logit garantissent $P \in [0, 1]$ et capturent la **saturation** aux extrêmes.

Résumé

Le Modèle de Probabilité Linéaire (MCO sur Y binaire) :

- ✓ Simple à estimer et interpréter
- ✓ Peut servir d'approximation locale
- × Hétéroscédasticité structurelle
- × Prédictions hors de $[0, 1]$
- × Effets marginaux constants (irréalistes)

Solutions préférées :

- **Probit** : $P(Y = 1|X) = \Phi(X'\beta)$
- **Logit** : $P(Y = 1|X) = \Lambda(X'\beta) = \frac{e^{X'\beta}}{1+e^{X'\beta}}$
- Nous pourrions, en principe, considérer d'autre distributions

$\Rightarrow$ Estimation par **maximum de vraisemblance**, propriétés asymptotiques bien établies.

Le cas d'une variable polytomique ordonnée

Définition

Variable qualitative à plus de deux modalités, ordonnées entre elles.

Exemple

- **Quantitative discrétisée** : Part des produits innovants dans le CA

  $$[0\% - 10\%],\ ]10\% - 30\%],\ ]30\% - 70\%],\ ]70\% - 100\%]$$

- **Appréciation subjective** : Importance de la R&D comme déterminant

  "Pas du tout", "Un peu", "Moyennement", "Beaucoup"

Remarque

Dans les deux cas, les modalités traduisent un ordre qui indique l'intensité de la variable.

Modèle latent

Variable latente continue

Le modèle latent représente la "vraie valeur" de la variable :

$$y_i^* = X_i b + u_i, \quad i = 1, \ldots, N$$

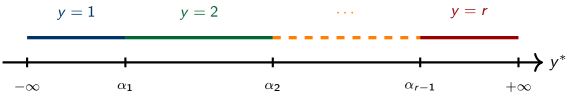
Observation par intervalle

La variable observable prend r modalités :

$$y_i = \begin{cases} 1 & \text{si } \alpha_0 < y_i^* \leq \alpha_1 \\ 2 & \text{si } \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & \\ r & \text{si } \alpha_{r-1} < y_i^* \leq \alpha_r \end{cases}$$

avec par convention $\alpha_0 = -\infty$ et $\alpha_r = +\infty$.

Représentation graphique



Seuils

- **Seuils connus** : cas des variables quantitatives discrétisées
- **Seuils inconnus** : cas des appréciations subjectives (à estimer)

Probabilités des modalités

Proposition

La probabilité d'observer la modalité j est :

$$\begin{aligned}\mathbb{P}(y_i = j) &= \mathbb{P}(\alpha_{j-1} < y_i^* \leq \alpha_j) \\ &= \mathbb{P}(y_i^* \leq \alpha_j) - \mathbb{P}(y_i^* \leq \alpha_{j-1})\end{aligned}$$

pour $j = 1, \ldots, r$.

Modèle Probit ordonné

Si $u_i \sim \mathcal{N}(0, \sigma^2)$:

$$\mathbb{P}(y_i = j) = \Phi\left(\frac{\alpha_j - X_i b}{\sigma}\right) - \Phi\left(\frac{\alpha_{j-1} - X_i b}{\sigma}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Le cas d'une variable de comptage

Définition

Variable prenant ses valeurs dans $\mathbb{N} = \{0, 1, 2, \ldots\}$, représentant le nombre d'occurrences d'un événement.

Exemple (Brevets)

Le nombre de brevets déposés par une entreprise sur une année :

- Valeurs entières positives ou nulles
- Événements relativement rares
- Beaucoup d'entreprises à 0 brevet

Remarque

Ce n'est pas une variable quantitative classique : elle ne peut pas prendre de valeurs négatives et a une nature discrète.

Modélisation

Espérance conditionnelle

L'espérance étant toujours strictement positive, on utilise une forme exponentielle :

$$\mathbb{E}(y_i \mid X_i, b) = \exp(X_i b + u_i) > 0$$

Loi de Poisson

On suppose que y_i suit une loi de Poisson de paramètre $\lambda_i = \exp(X_i b)$:

$$\mathbb{P}(y_i = k) = \frac{\exp(-\lambda_i)\lambda_i^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Sources d'aléa

Double source d'aléa

1. **Erreur sur la moyenne** : $\exp(u_i)$ (incertitude sur l'espérance)
2. **Tirage de Poisson** : aléa intrinsèque du processus de comptage

Distinction

- **Modèle de Poisson homogène** : $\mathbb{V}[\exp(u_i)] = 0$ (pas d'hétérogénéité inobservée)
- **Modèle de Poisson hétérogène** : $\mathbb{V}[\exp(u_i)] > 0$ (hétérogénéité présente)

Remarque

Le modèle de Poisson homogène correspond à une loi de durée exponentielle pour le temps entre événements.

SYNTHÈSE

Récapitulatif des deux parties

Partie I : Explicatives

Partie II : Expliquées

Synthèse —

Modèle	**Interprétation de b_j**	**Remarque**
Sans constante	$\mathbb{E}(y \mid \text{groupe } j)$	Moyenne du groupe
Avec constante	$\mathbb{E}(y \mid j) - \mathbb{E}(y \mid k)$	Écart à la référence k
Avec X	Idem, à X fixé	Effet contrôlé
Avec interactions	Effet hétérogène	Centrer X simplifie

Points clés — Variables qualitatives **à droite**

- Toujours indiquer la modalité de référence
- Le test de Fisher teste l'égalité des moyennes entre groupes
- Centrer les variables avant les produits croisés facilite l'interprétation

Synthèse — Partie II : Variables expliquées

Type	Valeurs	Modèle	Estimation
Dichotomique	$\{0, 1\}$	Logit/Probit	MV
Polytomique ordonnée	$\{1, \ldots, r\}$	Probit ordonné	MV
Comptage	$\mathbb{N}$	Poisson	MV/PMV
Censurée	Continue + sélection	Tobit/Heckman	MV

Principe commun — Variables qualitatives à **gauche**

1. Spécifier un **modèle latent** pour le phénomène sous-jacent
2. Définir le **lien** entre variable latente et variable observée
3. **Estimer** par maximum de vraisemblance

Pour aller plus loin

Chapitres suivants

- **Chapitre 2** : Maximum de vraisemblance (théorie et propriétés)
- **Chapitre 3** : Modèles Logit et Probit en détail
- **Chapitre 4** : Variables polytomiques
- **Chapitres 5** : Extensions

Références

-
- Gouriéroux, C. (1989). *Économétrie des variables qualitatives*
- Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*