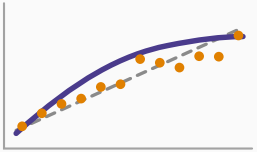
# Économétrie avancée

## Chapitre 7 : Variables qualitatives (indicatrices)



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

# Plan du chapitre

# Avant de commencer

# Pourquoi ce chapitre

**Pourquoi ce chapitre ?**

Le genre, la région, le secteur d'activité, un style architectural : autant de facteurs **qualitatifs** qui n'ont pas de magnitude numérique naturelle, mais qui influencent souvent $y$ de façon décisive. Ce chapitre montre comment les faire entrer dans le cadre de la régression multiple sans rien perdre de sa rigueur.

# Variables indicatrices simples

### Variable indicatrice (dummy)

Une **variable indicatrice** (ou *dummy*) code un fait binaire par $0$ ou $1$. Le choix de la modalité codée $1$ est arbitraire mais doit être **explicite** dans le nom de la variable (par exemple `female` plutôt que `gender`, pour éviter toute ambiguïté).

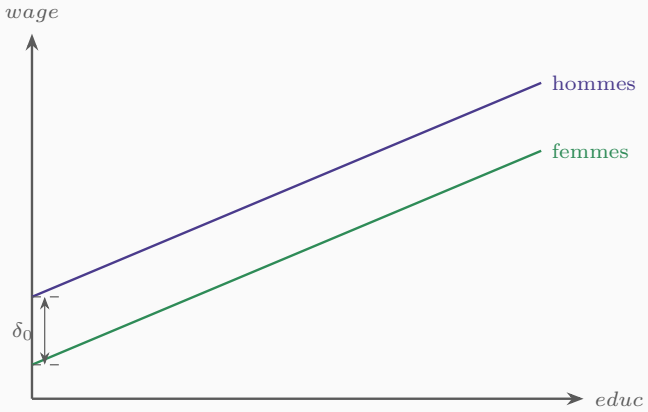
### Interprétation dans une régression

Dans $wage = \beta_0 + \delta_0\, female + \beta_1\, educ + u$, sous l'hypothèse habituelle $E(u \mid \cdot) = 0$,

$$\delta_0 = E(wage \mid female = 1, educ) - E(wage \mid female = 0, educ) :$$

c'est un **décalage d'ordonnée à l'origine** entre les deux groupes, à niveau d'éducation identique — les deux droites sont parallèles.

# Décrire l'information qualitative (suite)



# Décrire l'information qualitative (suite 2)

**Le piège de la variable indicatrice (dummy variable trap)**

Avec deux groupes seulement, il faut **une seule** indicatrice plus l'intercept — pas deux. Inclure à la fois `female` et `male` avec un intercept général crée une colinéarité parfaite ($female + male = 1$). Le groupe omis (ici les hommes) devient le **groupe de référence** ; tous les coefficients d'indicatrices se lisent comme des écarts à ce groupe.

**Écart salarial homme-femme**

$\widehat{wage} = -1.57 - 1.81\, female + .572\, educ + .025\, exper + .141\, tenure$ : à éducation, expérience et ancienneté égales, une femme gagne en moyenne 1,81 $ de moins par heure qu'un homme. Sans les variables de contrôle, l'écart brut est de 2,51 $ — la régression multiple isole la part de l'écart non expliquée par des différences de capital humain observées.

**Effet en pourcentage, exact et approché**

Si $\log(y) = \cdots + \delta_1 x_1 + u$ où $x_1$ est une indicatrice, l'approximation $100\,\delta_1$ donne l'écart en %; l'écart **exact** est $100[\exp(\delta_1) - 1]$, à utiliser dès que $|\delta_1|$ n'est pas petit (comme pour tout coefficient log-linéaire, chapitre 6).

# Catégories multiples et information ordinale

# Plus de deux groupes

**Règle générale : $g$ groupes, $g-1$ indicatrices**

Pour distinguer $g$ groupes, on inclut un intercept général et $g-1$ indicatrices (une par groupe, sauf le groupe de référence). Chaque coefficient mesure l'écart d'ordonnée à l'origine entre son groupe et le groupe de référence.

**Quatre profils croisés**

En croisant genre et statut marital (référence : hommes célibataires), $\widehat{\log(wage)} = .321 + .213\,marrmale - .198\,marrfem - .110\,singfem + .079\,educ + \ldots$ Les hommes mariés gagnent environ 21,3 % de plus que les hommes célibataires ; les femmes mariées, 19,8 % de moins. Pour tester directement l'écart entre deux groupes non-référence (ici, célibataires vs mariées), il faut soit recombiner les coefficients avec leur covariance, soit changer de groupe de référence et ré-estimer.

# Information ordinale

## Ordinal vs cardinal

Une variable **ordinale** (note de crédit, catégorie de beauté, tranche de classement) a un ordre mais pas d'écart interprétable entre modalités. L'entrer telle quelle impose un effet **constant** par unité, hypothèse souvent injustifiée. Solution : créer une indicatrice par modalité (sauf la modalité de référence), ce qui laisse l'effet libre de varier d'un échelon à l'autre.

## Application : classement des écoles et salaires

Découper le rang d'une école de commerce en tranches (`top10`, `r11_25`, ...) plutôt que de l'entrer en niveau améliore nettement l'ajustement ($\bar{R}^2$ passant de .836 à .905) : l'effet du classement sur le salaire de sortie n'est manifestement pas linéaire en rang. Ce principe s'applique à toute variable de notation — rating de crédit obligataire, scoring de risque — dès lors que rien ne garantit une relation linéaire avec la variable expliquée.

# Interactions impliquant des indicatrices

# Interaction entre deux indicatrices

**Retrouver les catégories croisées**

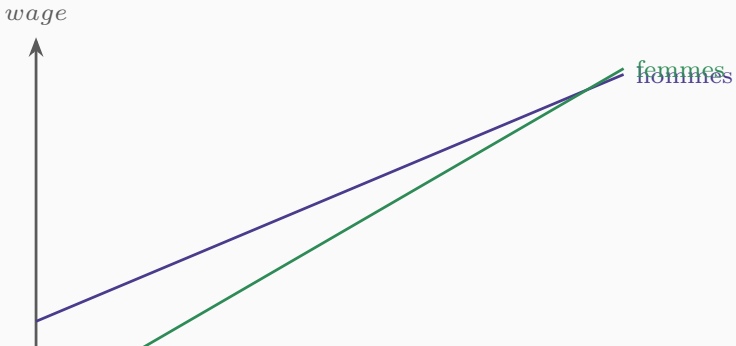
Ajouter $female \times married$ à un modèle contenant déjà `female` et `married` séparément permet à l'écart de mariage de différer entre hommes et femmes — c'est rigoureusement équivalent à définir directement les quatre indicatrices de groupe (chapitre précédent), mais l'écriture en interaction rend le test de l'hétérogénéité ($H_0$ : coefficient d'interaction $= 0$) immédiat.

# Pentes différentes selon le groupe

## Indicatrice × variable continue

$$\log(wage) = (\beta_0 + \delta_0\, female) + (\beta_1 + \delta_1\, female)\, educ + u$$

donne un intercept et une pente propres à chaque groupe. $\delta_0$ capture l'écart d'intercept, $\delta_1$ l'écart de **rendement de l'éducation** entre les groupes. Économétriquement, ce modèle s'estime en régressant $y$ sur $female$, $educ$ et $female \times educ$.



## Statistique de Chow

Pour tester si un même modèle décrit deux groupes ($H_0$ : mêmes coefficients dans les deux), on compare la $SSR$ de la régression **poolée** ($SSR_p$) à la somme des $SSR$ des régressions **séparées** par groupe ($SSR_1 + SSR_2$) :

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]/(k+1)}{(SSR_1 + SSR_2)/[n - 2(k+1)]} \sim F_{k+1,\, n-2(k+1)} \text{ sous } H_0.$$

Équivalent à tester conjointement la significativité de l'indicatrice de groupe et de toutes ses interactions avec les régresseurs.

## Chow global vs différence d'intercept seule

Le test de Chow interdit **toute** différence, y compris de pente. S'il rejette $H_0$, il faut identifier *quelle* différence compte : tester seulement les interactions ($H_0$ : pentes égales, intercepts libres) est souvent plus informatif — c'est ce qu'on a fait ci-dessus pour l'écart de rendement de l'éducation.

# Le modèle de probabilité linéaire

# Une variable dépendante binaire

## Modèle de probabilité linéaire (MPL)

Quand $y \in \{0, 1\}$, $P(y = 1 \mid x) = E(y \mid x)$. Le modèle

$$P(y = 1 \mid x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

s'estime par MCO comme n'importe quel modèle linéaire; $\hat{\beta}_j$ s'interprète comme la variation de la **probabilité de succès** pour une hausse d'une unité de $x_j$, toutes choses égales par ailleurs.

## Participation au marché du travail

$\widehat{inlf} = .586 - .0034\, nwifeinc + .038\, educ + .039\, exper - .0006\, exper^2 - .016\, age - .262\, kidslt6 + .013\, kidsge6$ : chaque enfant de moins de 6 ans réduit la probabilité de participation de 26,2 points, alors qu'un enfant plus âgé n'a quasiment aucun effet — un contraste économiquement très parlant.

# Une variable dépendante binaire (suite)

**Les limites structurelles du MPL**

Le MPL peut prédire des probabilités hors de $[0, 1]$, impose un effet marginal **constant** (irréaliste sur toute la plage de $x$), et souffre nécessairement d'hétéroscédasticité puisque $\mathrm{Var}(y \mid x) = p(x)[1 - p(x)]$ dépend de $x$. Cela ne biaise pas $\hat{\beta}_j$ (chapitre 3) mais invalide les erreurs-types usuelles — problème traité au chapitre 8. En pratique, le MPL reste très utilisé pour son interprétabilité directe, surtout près du centre de la distribution des régresseurs.

# Évaluation de politiques

# Groupes de traitement et de contrôle, auto-sélection

## Le problème de l'auto-sélection

Dans $y = \beta_0 + \beta_1 \, partic + u$, si la participation à un programme n'est pas assignée aléatoirement, $E(u \mid partic = 1) \neq E(u \mid partic = 0)$ est plausible : les participants diffèrent systématiquement des non-participants sur des facteurs **non observés** corrélés à $y$. La régression multiple atténue le problème en contrôlant les facteurs observables, mais ne le résout pas si des facteurs pertinents restent inobservés — situation qui appelle les méthodes plus avancées (variables instrumentales, données de panel) des chapitres suivants.

## Application : subventions à la formation et discrimination au crédit

Deux usages typiques du MPL en évaluation de politiques : tester la discrimination dans l'octroi de prêts ($approved = \beta_0 + \beta_1 \, nonwhite + \beta_2 \, income + ...$, où $\beta_1 < 0$ signale une discrimination *ceteris paribus*), et mesurer l'effet causal d'une subvention à la formation sur la productivité d'une firme. Dans les deux cas, la validité de l'interprétation causale dépend entièrement de la qualité des variables de contrôle incluses.

# Synthèse

# Ce qu'il faut retenir

## L'essentiel du chapitre 7

- Une indicatrice s'interprète comme un décalage d'intercept ; $g$ groupes exigent $g - 1$ indicatrices plus un intercept, sous peine de piège de colinéarité.
- L'information ordinale se code par indicatrices par modalité plutôt qu'en niveau, sauf si la linéarité est justifiée.
- Les interactions indicatrice $\times$ continue permettent des pentes différentes par groupe ; le test de Chow généralise ceci à une différence globale entre groupes.
- Le modèle de probabilité linéaire étend la régression MCO aux variables dépendantes binaires, avec des limites (bornes, hétéroscédasticité) à garder en tête.

## Vers le chapitre 8

Le MPL a mis en évidence un problème plus général : l'hétéroscédasticité, c'est-à-dire une variance de l'erreur qui dépend des régresseurs. Le chapitre suivant traite ce problème de façon systématique — comment le détecter, et comment corriger l'inférence.