Econométrie 2

Chapitre 3 : modèles binaires.

ENSAE 2021-2022

Michael Visser

CREST-ENSAE

Plan

Motivation

- On cherche à expliquer Y binaire par $X = (1, X_1, ..., X_{K-1})' \in \mathbb{R}^K$.
- Les deux valeurs possibles de Y étant arbitraires, on posera toujours $Y \in \{0, 1\}$.
- Les variables discrètes (et en particulier binaire) sont très largement répandues :
  - En microéconomie : activité vs inactivité, emploi vs chômage, consommation ou non d'un bien durable...
  - En risques de crédit : défaut ou non d'un emprunteur.
  - En assurance : sinistre ou non.
  - En biostatistique : individu guéri vs malade, traitement efficace ou non.
  - En sciences sociales : obtention d'un diplôme ou non, couple vs célibat, vote vs abstention etc.

Idées générales

- Les modèles linéaires sont mal adaptés pour étudier ce genre de variables...
- ... Mais on imposera un modèle linéaire sur une variable latente Y^* liée à Y.
- Comme les modèles correspondants sont non-linéaires en Y, nécessité de bien réfléchir aux paramètres d'intérêt.
- Le fait de ne pas observer Y^* impliquera la nécessité de normalisation et d'hypothèses paramétriques.
- Utilisation du maximum de vraisemblance (conditionnel aux X).
- Ces idées se retrouveront dans les chapitres suivants.

Plan

Modélisation

- Les modèles linéaires sont mal adaptés pour étudier ce genre de variables. En effet, si $Y \in \{0, 1\}$, on a

$$E(Y|X) = P(Y = 1|X) \in [0,\, 1]. \tag{1}$$

Dans un modèle linéaire (sous hypothèse d'exogénéité $E(\varepsilon|X) = 0$), on a $E(Y|X) = X'\beta_0$. Mais rien n'assure que $X'\beta_0 \in [0,\, 1]$.

- Pour que (1) soit satisfaite, on va supposer que

$$E(Y|X) = F(X'\beta_0), \tag{2}$$

où $F(.)$ est une fonction (connue) strictement croissante bijective de $\mathbb{R}$ dans $]0,\, 1[$, donc une fonction de répartition (fdr).

- N.B. : l'équation (2) est celle d'un modèle linéaire généralisé (GLM en anglais), c'est-à-dire un modèle de la forme :

$$h(E(Y|X)) = X'\beta_0,$$

où h est une fonction connue (dite fonction de lien).

Modèles non-linéaires et variables latentes

- Le modèle (2) peut s'interpréter en termes de *variables latentes*.
- Supposons qu'il existe une variable continue $Y^* \in \mathbb{R}$ telle que

$$Y = \mathbb{1}\{Y^* \geq 0\}.$$

- Supposons par ailleurs que Y^* suive un modèle linéaire :

$$Y^* = X'\beta_0 + \varepsilon, \tag{3}$$

  où $-\varepsilon$ est indépendante de X et a pour fonction de répartition F. Alors

$$P(Y = 1|X) = P(X'\beta_0 + \varepsilon \geq 0|X) = P(-\varepsilon \leq X'\beta_0|X) = F(X'\beta_0).$$

- On retrouve donc l'équation (2).
- L'interprétation en termes de variables latentes est, très souvent, naturelle.

Exemple 1 : microéconomie.

- Y =choix binaire de la part d'un agent. Soit U_1 l'utilité (espérée) de l'agent s'il décide $Y = 1$, U_0 son utilité s'il décide $Y = 0$.
- Posons alors $Y^* = U_1 - U_0$ la différence d'utilité entre les deux choix.
- Si l'agent est rationnel, il décide en maximisant son utilité espérée :

$$Y = \mathbb{1}\{U_1 \geq U_0\} = \mathbb{1}\{Y^* \geq 0\}.$$

Exemple 2 : finance d'entreprise.

- Le défaut ($Y = 1$) d'une entreprise survient lorsque la dette de l'entreprise D dépasse un certain seuil S (éventuellement aléatoire).
- On a alors $Y^* = D - S$.

Interprétation en termes de variables latentes

Exemple 3 : biostatistique.

- Y =1 si malade, 0 sinon.
- Un individu est guéri lorsque le nombre de bactéries N (par exemple) est descendu en dessous d'un certain seuil S, éventuellement fonction de l'individu.
- On a alors $Y^* = N - S$.

Exemple 4 : éducation.

- $Y = 1$ si l'individu obtient son diplôme, 0 sinon.
- On obtient son diplôme si notre moyenne M est supérieure à un seuil fixé s.
- On a alors $Y^* = M - s$.

Deux exemples importants : le probit et le logit

- A priori n'importe quel choix est possible pour F.
- Les plus courants sont
  - $F = \Phi$, fdr d'une $\mathcal{N}(0,1)$: modèle probit ;
  - $F(x) = \Lambda(x) = 1/(1+\exp(-x))$, fdr d'une loi *logistique* : modèle logit.
- La différence entre les deux est assez faible. Lorsque $|x| \to +\infty$, $\varphi(x) = \Phi'(x) \propto e^{-x^2/2}$ et $\Lambda'(x) = \Lambda(x)(1-\Lambda(x)) = O(e^{-|x|})$.

$\Rightarrow$ queues plus épaisses pour la loi logistique.

Paramètres et effets marginaux

- Qualitativement, la $j-$ème composante X_j de X aura un effet positif sur $P(Y=1|X)$ ssi $\beta_{0j}>0$.
- Quantitativement, l'interprétation de β_{0j} est plus délicate.
- Dans le modèle linéaire standard $E(Y|X)=X'\beta_0$, le paramètre β_{0j} de X_j peut s'interpréter comme l'effet marginal de X_j :

$$\frac{\partial E(Y|X_1=x_1,...,X_{K-1}=x_{K-1})}{\partial x_j}=\beta_{0j}.$$

- Cette valeur est indépendante de $x=(x_1,...,x_{K-1})$.

Paramètres et effets marginaux

- Mais dans les modèles binaires (et non-linéaires plus généralement), l'effet marginal de X_j n'est plus β_{0j}, et il dépend de x :

$$\frac{\partial E(Y|X=x)}{\partial x_j} = \left.\frac{\partial F(u)}{\partial u}\right|_{u=x'\beta_0} \frac{\partial x'\beta_0}{\partial x_j} = f(x'\beta_0)\beta_{0j} \quad \text{avec } f = F'.$$

- Remarque 1 : si f est symétrique unimodale, l'effet d'une variable sur $P(Y=1|X)$ est d'autant plus fort que $x'\beta_0$ est proche de 0, soit $P(Y=1|X) \simeq 0.5$.

- Remarque 2 : on a toujours

$$\frac{\beta_{0l}}{\beta_{0j}} = \frac{\partial E(Y|X=x)/\partial x_l}{\partial E(Y|X=x)/\partial x_j}$$

Donc la comparaison des différents paramètres est licite.

Paramètres et effets marginaux

- Outre l'estimation de β_{0j}, il est intéressant d'estimer l'effet marginal moyen :
$$\Delta_j = E\left[f(X'\beta_0)\right]\beta_{0j}.$$
C'est l'effet sur Y d'une augmentation marginale de X_j appliquée à tous les individus.

- On peut également se concentrer sur des sous-populations, en calculant l'effet marginal moyen pour les individus vérifiant $X \in A$ (par exemple) : $E\left[f(X'\beta_0)|X \in A\right]\beta_{0j}$ ou encore l'effet marginal à la moyenne, $f\left(E(X)'\beta_0\right)\beta_{0j}$.

- Pour les variables explicatives discrètes (dichotomiques), l'effet marginal est remplacé par
$$F(x'_{-j}\beta_{0-j} + \beta_{0j}) - F(x'_{-j}\beta_{0-j}).$$
où $x_{-j} = (1, x_1, ..., x_{j-1}, x_{j+1}, ..., x_{K-1})'$. L'effet moyen d'un changement universel de X_j de 0 à 1 est alors :
$$E\left[F(X'_{-j}\beta_{0-j} + \beta_{0j}) - F(X'_{-j}\beta_{0-j})\right].$$

Une spécificité du modèle logit : les odds-ratios.

- On définit le risque (ou odd) comme égal à $P(Y=1|X)/P(Y=0|X)$.
- Dans le cas du logit :

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{1/(1+e^{-x'\beta_0})}{e^{-x'\beta_0}/(1+e^{-x'\beta_0})} = e^{x'\beta_0}$$

- Considérons une variable explicative $X_j \in \{0,1\}$. On a alors :

$$e^{\beta_{0j}} = \frac{P(Y=1|X_{-j}=x_{-j}, X_j=1)/P(Y=0|X_{-j}=x_{-j}, X_j=1)}{P(Y=1|X_{-j}=x_{-j}, X_j=0)/P(Y=0|X_{-j}=x_{-j}, X_j=0)}.$$

- Donc $e^{\beta_{0j}}$ est égal au rapport des risques (odds-ratio) correspondant à $X_j=1$ et $X_j=0$. Il est indépendant de la valeur de X_{-j}.

Plan

Identification

- Revenons à l'équation :

$$Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\}.$$

- Deux questions : (i) pourquoi fixer le seuil à 0 ? (ii) pourquoi fixer la variance de ε (à 1 pour le probit, à $\pi^2/3$ pour le logit) ?
- Raison : le modèle n'est pas identifiable sinon. En effet, on a :

$$Y = \mathbb{1}\{\beta_{01} + X'_{-1}\beta_{0-1} + \varepsilon \geq s\} \iff Y = \mathbb{1}\{\beta_{01} - s + X'_{-1}\beta_{0-1} + \varepsilon \geq 0\}.$$

- En d'autres termes, on ne peut pas identifier séparément la constante β_{01} et le seuil s. On fixe donc (arbitrairement) s à 0.
- De même, on ne peut pas identifier de façon jointe β_0 et la variance σ_0^2 du résidu ε. En effet,

$$Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\} \iff Y = \mathbb{1}\{X'(\beta_0/\sigma_0) + \varepsilon/\sigma_0 \geq 0\}.$$

- On fixe donc arbitrairement σ_0.

Proposition 1

Si s et σ_0 sont fixés et $E(XX')$ est inversible, le modèle est identifié.

Preuve : soit P_β la loi des observations lorsque le vrai paramètre est β. Il s'agit de montrer que la fonction $\beta \mapsto P_\beta$ est injective. Dans notre modèle conditionnel, on peut montrer que l'identification est équivalente à

$$P_\beta(Y=1|X) = P_{\beta'}(Y=1|X) \Rightarrow \beta = \beta' \quad \forall (\beta, \beta').$$

Or

$$(E(XX') \text{ est inversible }) \Longleftrightarrow (X'\lambda = 0 \Longrightarrow \lambda = 0)$$

Par conséquent,

$$\begin{aligned} P_\beta(Y=1|X) = P_{\beta'}(Y=1|X) \quad &\Longleftrightarrow \quad F(X'\beta) = F(X'\beta') \\ &\Longleftrightarrow \quad X'\beta = X'\beta' \\ &\Longleftrightarrow \quad \beta = \beta' \ \square \end{aligned}$$

Estimation du modèle : le maximum de vraisemblance

- On s'intéresse maintenant à l'estimation de β_0 à partir d'un échantillon i.i.d. $((Y_1, X_1), ..., (Y_n, X_n))$.
- Comme le modèle est paramétrique, on peut l'estimer par maximum de vraisemblance.
- La densité de Y conditionnellement à X s'écrit :

$$
\begin{aligned}
P(Y = y|X = x) &= [P(Y = 1|X = x)]^y \; [P(Y = 0|X = x)]^{1-y} \\
&= F(x'\beta)^y \; (1 - F(x'\beta))^{1-y}.
\end{aligned}
$$

- La fonction de vraisemblance d'un échantillon i.i.d. $(\mathrm{Y}, \mathrm{X}) = ((Y_1, X_1), ..., (Y_n, X_n))$ conditionnellement à X s'écrit alors

$$
\mathcal{L}_n(\mathrm{Y}|\mathrm{X}; \beta) = \prod_{i=1}^{n} F(X_i'\beta)^{Y_i} \; (1 - F(X_i'\beta))^{1-Y_i}.
$$

- Un estimateur du maximum de vraisemblance est alors défini par :

$$\widehat{\beta} \in \arg\max_{\beta \in \mathbb{R}^K} \mathcal{L}_n(\mathrm{Y}|\mathrm{X};\beta).$$

- Remarquez que cet estimateur n'est pas forcément unique, et il peut ne pas exister.

- Remarquez aussi que $\mathcal{L}_n(\mathrm{Y}|\mathrm{X};\beta)$ est la vraisemblance *conditionnelle* à X. En notant $g(X_i)$ la densité de X_i, la vraisemblance non-conditionnelle s'écrit

$$\mathcal{L}_n(\mathrm{Y},\mathrm{X};\beta) = \prod_{i=1}^{n} F(X_i'\beta)^{Y_i}\,(1 - F(X_i'\beta))^{1-Y_i} g(X_i).$$

  Comme dans la pratique on ne s'intéresse pas à la distribution de X, on se focalise sur la vraisemblance conditionnelle.

Conditions du premier ordre du programme

- On maximise plutôt la log-vraisemblance :

$$\ell_n(\mathrm{Y}|\mathrm{X};\beta) = \sum_{i=1}^{n} Y_i \ln\left(F(X_i'\beta)\right) + (1 - Y_i)\ln\left(1 - F(X_i'\beta)\right)$$

- On a $\partial F(X_i'\beta)/\partial\beta = f(X_i'\beta)X_i$. Donc :

$$\frac{\partial \ell_n}{\partial \beta}(\mathrm{Y}|\mathrm{X};\beta) = \sum_{i=1}^{n}\left[Y_i \frac{f(X_i'\beta)}{F(X_i'\beta)} + (1 - Y_i)\frac{-f(X_i'\beta)}{1 - F(X_i'\beta)}\right] X_i$$

  Soit encore

$$\frac{\partial \ell_n}{\partial \beta}(\mathrm{Y}|\mathrm{X};\beta) = \sum_{i=1}^{n} \frac{f(X_i'\beta)}{F(X_i'\beta)(1 - F(X_i'\beta))}\left[Y_i - F(X_i'\beta)\right] X_i. \qquad (4)$$

- Les conditions du premier ordre s'écrivent donc :

$$\sum_{i=1}^{n} \frac{f(X_i'\widehat{\beta})}{F(X_i'\widehat{\beta})(1 - F(X_i'\widehat{\beta}))}\left[Y_i - F(X_i'\widehat{\beta})\right] X_i = 0 \qquad (5)$$

  qui n'admet pas de solution analytique simple en général.

Existence et unicité de solution

- Si une variable dichotomique X_j est telle que : si $x_{ij} = 1 \Rightarrow y_i = 1$ pour tout i (ou $x_{ij} = 1 \Rightarrow y_i = 0$ pour tout i), l'estimateur n'existe pas.
- En effet, $\partial \ell_n / \partial \beta_j$ s'écrit alors (on considère le cas $x_{ij} = 1 \Rightarrow y_i = 1$ pour tout i)

$$\sum_{i=1}^{n} \frac{f(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))} [y_i - F(x_i'\beta)]\, x_{ij} = \sum_{i:x_{ij}=1} \frac{f(x_i'\beta)}{F(x_i'\beta)} > 0 \quad \forall \beta$$

- L'exemple montre que, parmi les observations telles que $x_{ij} = 1$, la variable y_i doit varier à travers i pour pouvoir estimer β_{0j}. En absence de variation, des logiciels comme Stata indiquent que le paramètre β_{0j} n'est pas identifiable et "expulse" automatiquement la variable x_j du modèle.

Existence et unicité de solution

- Dans le cas du modèle logit, on a $\Lambda' = \Lambda(1-\Lambda)$, donc

$$\frac{\partial^2 \ell_n}{\partial\beta\partial\beta'}(\mathrm{Y}|\mathrm{X};\beta) = -\sum_{i=1}^{n} \Lambda'(X_i'\beta)X_iX_i' << 0.$$

  La matrice des dérivés secondes est définie négative. La log-vraisemblance est donc strictement concave $\Rightarrow$ les conditions de premier ordre ont au plus une solution, et cette solution correspond au maximum global.

- Dans le cas du modèle probit, on peut également montrer que la log-vraisemblance est bien strictement concave.

- Dans le cas général, le programme n'est pas nécessairement concave et il peut y avoir plusieurs solutions. Il faut alors (dans l'idéal) vérifier que la solution corresponde bien à un maximum global.

Remarques sur l'optimisation

- Contrairement à l'estimateur des MCO, l'estimateur du maximum de vraisemblance ne peut pas, en général, s'exprimer d'une manière explicte.

- L'estimateur peut être obtenu numériquement par un algorithme de Newton-Raphson (il existe d'autres algorithmes). Partant de $\beta^{(0)}$ quelconque, on définit la suite $\left(\beta^{(m)}\right)_{m\in\mathbb{N}}$ par :

$$\beta^{(m+1)} = \beta^{(m)} - \left[\frac{\partial^2 \ell_n}{\partial\beta\partial\beta'}(\mathrm{Y}|\mathrm{X};\beta^{(m)})\right]^{-1} \frac{\partial \ell_n}{\partial\beta}(\mathrm{Y}|\mathrm{X};\beta^{(m)})$$

- Par concavité stricte de $\ell_n(\mathrm{Y}|\mathrm{X};\beta)$, la suite $\beta^{(m)}$, si elle converge, tend nécessairement vers l'EMV.

- Dans le cas d'un modèle logit ou probit les itérations convergent typiquement vite et rapidement.

Proposition 2

Sous plusieurs conditions techniques (voir cours de Statistique 1), on a $\widehat{\beta} \xrightarrow{P} \beta_0$ *et*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_1^{-1}(\beta_0)\right),$$

où $\mathcal{I}_1(\beta_0)$ *est l'information de Fisher associée à une observation. De plus,*

$$\mathcal{I}_1(\beta_0) = E\left(\frac{f^2(X'\beta_0)}{F(X'\beta_0)(1 - F(X'\beta_0))}XX'\right).$$

On peut l'estimer de façon convergente par :

$$\widehat{\mathcal{I}_1(\beta_0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{f^2(X_i'\widehat{\beta})}{F(X_i'\widehat{\beta})(1 - F(X_i'\widehat{\beta}))}X_iX_i'.$$

Rappel : l'EMV est asymptotiquement le meilleur estimateur "régulier" : si un autre estimateur $\widetilde{\beta}$ vérifie $\sqrt{n}\left(\widetilde{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, V)$, alors $V >> \mathcal{I}_1^{-1}(\beta_0)$.

Propriétés asymptotiques

Preuve de la formule sur $\mathcal{I}_1(\beta_0)$: on a

$$\mathcal{I}_1(\beta_0) = V\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\right)$$

où $\ell_1(Y|X;\beta_0)$ est la log-vraisemblance (évaluée en β_0) d'une observation :

$$\ell_1(Y|X;\beta_0) = Y \ln\left(F(X'\beta_0)\right) + (1-Y)\ln\left(1 - F(X'\beta_0)\right)$$

Par décomposition de la variance :

$$\mathcal{I}_1(\beta_0) = E\left[V\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\middle| X\right)\right] + V\left[E\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\middle| X\right)\right].$$

Or (cf. équation (4))

$$\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0) = \frac{f(X'\beta_0)}{F(X'\beta_0)(1-F(X'\beta_0))}\left[Y - F(X'\beta_0)\right] X.$$

Donc

$$E\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\middle| X\right) = 0$$

car $E\left(Y - F(X'\beta_0)|X\right) = 0$, et

$$\begin{aligned} V\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\middle| X\right) &= E\left[\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\right)\left(\frac{\partial \ell_1}{\partial \beta}(Y|X;\beta_0)\right)'\middle| X\right] \\ &= E\left[\frac{f(X'\beta_0)^2}{F(X'\beta_0)^2(1-F(X'\beta_0))^2}\left[Y - F(X'\beta_0)\right]^2 XX'\middle| X\right] \\ &= \frac{f^2(X'\beta_0)XX'}{F(X'\beta_0)(1-F(X'\beta_0))} \end{aligned}$$

car $E\left[(Y - F(X'\beta_0))^2\middle| X\right] = F(X'\beta_0)(1 - F(X'\beta_0))$. D'où le résultat.

Tests d'hypothèses.

- On souhaite tester une hypothèse du type

$$H_0 \ : R\beta_0 = 0 \ \text{ contre } \ H_1 \ : R\beta_0 \neq 0 \quad (R \ \text{ matrice } p \times K, \ p \leq K).$$

- Par exemple, $\beta_{0j} = 0$ ou $\beta_{02} = ... = \beta_{0K-1} = 0$ (c'est à dire $\beta_{0-1} = 0$).
- On utilise l'un des trois tests liés au maximum de vraisemblance : le test de Wald, le test du score ou le test de rapport de vraisemblance. Les statistiques de test correspondantes s'écrivent :

$$\begin{aligned}
\xi_n^W &= n\widehat{\beta}' R' \left[R \, \widehat{\mathcal{I}_1(\beta_0)}^{-1} \, R' \right]^{-1} R\widehat{\beta} \\
\xi_n^S &= \frac{1}{n} \frac{\partial \ell_n}{\partial \beta'}(\mathrm{Y}|\mathrm{X}; \widehat{\beta}_C) \, \widehat{\mathcal{I}_1(\beta_0)}^{-1} \, \frac{\partial \ell_n}{\partial \beta}(\mathrm{Y}|\mathrm{X}; \widehat{\beta}_C) \\
\xi_n^R &= 2 \left[\ell_n(\mathrm{Y}|\mathrm{X}; \widehat{\beta}) - \ell_n(\mathrm{Y}|\mathrm{X}; \widehat{\beta}_C) \right]
\end{aligned}$$

où $\widehat{\beta}_C$ est l'estimateur du maximum de vraisemblance contraint, i.e. estimé sous H_0.

Tests d'hypothèses.

- Concernant la statistique ξ_n^W, $\widehat{\mathcal{I}_1(\beta_0)}$ correspond à la formule de la page 24.
- Concernant la statistique ξ_n^S, il s'agit de cette même formule sauf que $\widehat{\beta}$ est remplacé par $\widehat{\beta}_C$.
- Remarquez que les trois statistiques ont tendance à être "petites" sous l'hypothèse H_0.
- Sous H_0, $\xi_n^T \xrightarrow{d} \chi^2(p)$ ($T = W, S$ ou R).
- La région critique d'un test de niveau asymptotique α est donc de la forme $\{\xi_n^T > q_p^{Chi-2}(1-\alpha)\}$ où $q_p^{Chi-2}(y)$ est le quantile d'ordre y d'un $\chi^2(p)$.
- Lorsqu'on teste $H_0 \ : \beta_{0j} = 0$ contre $H_1 \ : \beta_{0j} \neq 0$, le t-test habituel est le plus souvent utilisé. Ce test donne le même résultat que le test de Wald, car (à vérifier) $\xi_n^W = \left(\widehat{\beta}_j / se(\widehat{\beta}_j)\right)^2 \equiv t_j^2$ et $|t_j| > q^N(1-\alpha/2) \Leftrightarrow \xi_n^W > q_p^{Chi-2}(1-\alpha)$ où $q^N(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$.

Plan

Pouvoir explicatif du modèle

- On définit, de façon similaire au R^2, le pseudo-R^2 par :

$$\text{pseudo-}R^2 = 1 - \frac{\ell_n(\mathrm{Y}|\mathrm{X};\widehat{\beta})}{\ell_n(\mathrm{Y}|\mathrm{X};\widehat{\beta}_C)}$$

où $\widehat{\beta}_C$ est le paramètre estimé sous l'hypothèse nulle $\beta_{0-1} = 0$.

- Puisque $0 > \ell_n(\mathrm{Y}|\mathrm{X};\widehat{\beta}) \geq \ell_n(\mathrm{Y}|\mathrm{X};\widehat{\beta}_C)$, le pseudo-$R^2$ appartient à $]0,1]$. Il est proche de 1 lorsque

$$Y_i = 1 \text{ et } F(X_i'\widehat{\beta}) \simeq 1 \;\text{ ou }\; Y_i = 0 \text{ et } F(X_i'\widehat{\beta}) \simeq 0.$$

- Comme le R^2, le pseudo-R^2 augmente mécaniquement avec le nombre de variables.
- Autres indicateurs : table de concordance, score, pourcentage de paires concordantes...

Choix des variables

- Arbitrage entre :
  - l'accroissement du pouvoir explicatif du modèle ;
  - la perte de précision liée à l'estimation de nombreux paramètres.
- on peut faire des tests de nullité des variables, éventuellement via des procédures séquentielles (forward, backward, ...).
- Inconvénient : lorsque n tend vers l'infini, on est conduit à accepter la plupart des variables explicatives.
- on peut utiliser les critères d'information AIC (*Akaike Information Criterion*, Akaike, 1973) ou BIC (*Bayesian Information Criterion*, Schwarz, 1978).
- Ces critères sont utilisés pour résoudre le problème du choix de modèles. Supposons que l'on ait J modèles paramétriques possibles :

$$\left\{ \left(P_{\beta^{(1)}}\right)_{\beta^{(1)} \in B^{(1)}}, ..., \left(P_{\beta^{(J)}}\right)_{\beta^{(J)} \in B^{(J)}} \right\}.$$

  On souhaite sélectionner le vrai modèle.

Choix des variables

- Critère d'Akaike pour le modèle j ayant p_j paramètres :

$$\mathrm{AIC}(j) = \ell_n(\mathrm{Y}|\mathrm{X}; \widehat{\beta}^{(j)}) - p_j$$

  On choisit alors le modèle $j_0 = \arg\max_j \mathrm{AIC}(j)$.

- Ce critère ne conduit pas au bon choix lorsque n tend vers l'infini. En effet, le critère ne pénalise pas assez le nombre de paramètres.
- Pour corriger cela, Schwarz (1978) propose le critère suivant :

$$\mathrm{BIC}(j) = \ell_n(\mathrm{Y}|\mathrm{X}; \widehat{\beta}^{(j)}) - \frac{p_j}{2}\ln(n)$$

Plan

Intérêt du modèle linéaire

- Parfois, pour des raisons de simplicité, on estime un modèle de probabilité linéaire plutôt qu'un modèle logit ou probit :

$$E(Y|X) = X'\beta_0.$$

- Exemple : données de panel. Supposons que

$$E(Y_{it}|X_{it}, \alpha_i) = X_{it}'\beta_0 + \alpha_i,$$

où α_i est un effet individuel a priori corrélé aux X_{it}.

- On peut éliminer l'effet inobservé par différence ou par within :

$$E(Y_{it} - Y_{it-1}|X_{it}, X_{it-1}) = (X_{it} - X_{it-1})'\beta_0.$$

- Dans les modèles non-linéaires, ce n'est pas aussi simple car

$$E(Y_{it} - Y_{it-1}|X_{it}, X_{it-1}, \alpha_i) = F(X_{it}'\beta + \alpha_i) - F(X_{it-1}'\beta + \alpha_i).$$

- Par ailleurs, l'estimation par maximum de vraisemblance de (β, $\alpha_1, ..., \alpha_n$) n'est pas convergente à cause du problème des *paramètres incidents* : le nombre de paramètres tend vers l'infini avec n.

Modélisation et estimation

- Le modèle de probabilité linéaire peut se réécrire $Y = X'\beta_0 + \varepsilon$, avec

$$\varepsilon = \left| \begin{array}{lll} 1 - X'\beta_0 & \text{avec la probabilité (conditionnelle)} & X'\beta_0 \\ -X'\beta_0 & \text{avec une probabilité} & 1 - X'\beta_0 \end{array} \right.$$

- On a donc :

$$\begin{aligned} V(\varepsilon|X) &= E(\varepsilon^2|X) = X'\beta_0(1 - X'\beta_0)^2 + (1 - X'\beta_0)(X'\beta_0)^2 \\ &= X'\beta_0(1 - X'\beta_0). \end{aligned}$$

- Le modèle est hétéroscédastique. On peut l'estimer par MCO mais aussi par MCQG :
  - On estime β_0 par MCO : $\widehat{\beta}_{MCO}$.
  - On réestime β_0 par

$$\widehat{\beta}_{MCGQ} = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{X_i'\widehat{\beta}_{MCO}(1 - X_i'\widehat{\beta}_{MCO})} \left[Y_i - X_i'\beta\right]^2$$

- L'estimateur des MCQG est plus précis en théorie mais pas forcément en pratique, si $X_i'\widehat{\beta}_{MCO} \simeq 0$ ou $\simeq 1$ pour certains i.

Comparaison logit/probit/modèle linéaire

- La différence avec les modèles logit, probit et linéaire est que l'on ne choisit pas le même $F(.)$ dans $E(Y|X) = F(X'\beta_0)$: $F = \Lambda, \Phi$, ou Identité suivant le modèle considéré.
- Il existe des modèles semi-paramétriques où l'on suppose que $P(Y = 1|X) = F(X'\beta_0)$ avec F et β_0 inconnus. Équivaut à considérer le modèle latent $Y^* = X'\beta_0 + \varepsilon$ avec $\varepsilon \perp\!\!\!\perp X$ de loi inconnue.
- Ces modèles sont moins restrictifs mais plus difficiles à estimer.
- Par ailleurs, les résultats des logit, probit et modèles linéaires sont souvent très proches en termes d'effets marginaux.
- En termes de coefficients, on aura en général

$$\widehat{\beta}_{\text{logit}} \simeq 1.6\widehat{\beta}_{\text{probit}} \simeq 4\widehat{\beta}_{\text{linéaire}}.$$

Plan

Exemple : activité des femmes.

- On cherche à expliquer l'activité ($Y = 1$, $Y = 0$ sinon) des femmes suivant leur âge, leur diplôme et leur situation familiale (en couple ou non, nombre d'enfants de moins de 3 ans).
- On utilise les données de l'enquête emploi française 2012. On se concentre sur les femmes de moins de 65 ans ayant terminé leurs études.
- Modalités de la variable de diplôme DDIPL :
  1 Diplôme supérieur
  3 Baccalauréat + 2 ans
  4 Baccalauréat ou brevet professionnel ou autre diplôme de ce niveau
  5 CAP, BEP ou autre diplôme de ce niveau
  6 Brevet des collèges
  7 Aucun diplôme ou CEP

Code Stata

```
destring age ddipl, replace

* Sélection de l'échantillon : femme en 1ère interrogation, ayant fini leur
* études et de moins de 65 ans
keep if rgi=="1" & sexe=="2" & fordat!="" & acteu!="" & age<=65

gen en_couple = (TYPMEN5=="3" | TYPMEN5=="4")
gen actif = (acteu!="3")

logit actif c.age c.age#c.age NBENF3 en_couple i.ddipl

* Calcul des effets marginaux
margins, dydx(_all)
margins, dydx(_all) atmeans

* Calcul des odds ratios
logistic actif c.age c.age#c.age NBENF3 en_couple i.ddipl

* Comparaison des effets marginaux avec le probit
probit actif c.age c.age#c.age NBENF3 en_couple i.ddipl
margins, dydx(_all)
```

Résultats : coefficients du logit

```
Iteration 0:   log likelihood = -17112.651
Iteration 1:   log likelihood = -13273.506
Iteration 2:   log likelihood = -13110.555
Iteration 3:   log likelihood = -13110.084
Iteration 4:   log likelihood = -13110.084

Logistic regression                             Number of obs     =     29,248
                                                LR chi2(9)        =    8005.14
                                                Prob > chi2       =     0.0000
Log likelihood = -13110.084                     Pseudo R2         =     0.2339
```

actif	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.3588415	.008609	41.68	0.000	.3419681	.3757148
c.age#c.age	-.004898	.0000999	-49.03	0.000	-.0050938	-.0047022
NBENF3	-1.433909	.0459917	-31.18	0.000	-1.524051	-1.343766
en_couple	-.2182632	.0347124	-6.29	0.000	-.2862983	-.1502281
ddipl						
3	-.1536238	.0649168	-2.37	0.018	-.2808584	-.0263893
4	-.4915752	.0593656	-8.28	0.000	-.6079296	-.3752208
5	-.6802183	.0566461	-12.01	0.000	-.7912426	-.569194
6	-.890034	.0680838	-13.07	0.000	-1.023476	-.7565922
7	-1.411182	.0559724	-25.21	0.000	-1.520886	-1.301478
_cons	-3.342729	.180223	-18.55	0.000	-3.69596	-2.989499

Résultats : effets marginaux moyens

```
Average marginal effects                        Number of obs     =     29,248
Model VCE    : OIM

Expression   : Pr(actif), predict()
dy/dx w.r.t. : age NBENF3 en_couple 3.ddipl 4.ddipl 5.ddipl 6.ddipl 7.ddipl
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0129698	.0001931	-67.18	0.000	-.0133482	-.0125913
NBENF3	-.2050776	.0062575	-32.77	0.000	-.217342	-.1928131
en_couple	-.031216	.0049551	-6.30	0.000	-.0409278	-.0215042
ddipl						
3	-.0185428	.0078345	-2.37	0.018	-.0338981	-.0031874
4	-.0637538	.0075632	-8.43	0.000	-.0785773	-.0489302
5	-.0916208	.0073535	-12.46	0.000	-.1060334	-.0772083
6	-.1247672	.0097925	-12.74	0.000	-.1439602	-.1055741
7	-.2160232	.0080491	-26.84	0.000	-.2317991	-.2002474

Note: dy/dx for factor levels is the discrete change from the base level.

Résultats : effets marginaux à la moyenne

```
Conditional marginal effects                    Number of obs     =     29,248
Model VCE    : OIM

Expression   : Pr(actif), predict()
dy/dx w.r.t. : age NBENF3 en_couple 3.ddipl 4.ddipl 5.ddipl 6.ddipl 7.ddipl
at           : age             =    43.60835 (mean)
               NBENF3          =    .1043832 (mean)
               en_couple       =    .7026805 (mean)
               1.ddipl         =    .1625068 (mean)
               3.ddipl         =    .1498222 (mean)
               4.ddipl         =    .1864401 (mean)
               5.ddipl         =     .218032 (mean)
               6.ddipl         =    .0816124 (mean)
               7.ddipl         =    .2015864 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0068105	.0001604	-42.46	0.000	-.0071248	-.0064961
NBENF3	-.1428768	.0049677	-28.76	0.000	-.1526133	-.1331404
en_couple	-.0217481	.0034523	-6.30	0.000	-.0285144	-.0149817
ddipl						
3	-.0097966	.0041521	-2.36	0.018	-.0179346	-.0016587
4	-.036382	.0043838	-8.30	0.000	-.0449742	-.0277898
5	-.054696	.0044551	-12.28	0.000	-.0634279	-.0459642
6	-.0784038	.0067935	-11.54	0.000	-.0917188	-.0650888
7	-.1543896	.006309	-24.47	0.000	-.1667551	-.1420242

Note: dy/dx for factor levels is the discrete change from the base level.

Résultats : rapports des risques (odds-ratios)

```
Logistic regression                             Number of obs     =     29,248
                                                LR chi2(9)        =    8005.14
                                                Prob > chi2       =     0.0000
Log likelihood = -13110.084                     Pseudo R2         =     0.2339
```

actif	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.43167	.0123253	41.68	0.000	1.407715	1.456032
c.age#c.age	.9951139	.0000994	-49.03	0.000	.9949191	.9953088
NBENF3	.2383754	.0109633	-31.18	0.000	.2178277	.2608613
en_couple	.8039138	.0279058	-6.29	0.000	.7510386	.8605116
ddipl						
3	.8575946	.0556723	-2.37	0.018	.7551353	.9739559
4	.6116621	.0363117	-8.28	0.000	.544477	.6871375
5	.5065064	.0286916	-12.01	0.000	.4532812	.5659815
6	.4106418	.0279581	-13.07	0.000	.3593438	.4692629
7	.2438549	.0136491	-25.21	0.000	.2185182	.2721292
_cons	.0353404	.0063691	-18.55	0.000	.0248236	.0503127

Résultats : sortie du probit

```
Iteration 0:   log likelihood = -17112.651
Iteration 1:   log likelihood = -13167.282
Iteration 2:   log likelihood = -13144.263
Iteration 3:   log likelihood = -13144.235
Iteration 4:   log likelihood = -13144.235

Probit regression                               Number of obs     =     29,248
                                                LR chi2(9)        =    7936.83
                                                Prob > chi2       =     0.0000
Log likelihood = -13144.235                     Pseudo R2         =     0.2319
```

actif	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.208209	.0049031	42.46	0.000	.1985991	.2178189
c.age#c.age	-.0028442	.0000567	-50.19	0.000	-.0029553	-.0027331
NBENF3	-.810221	.0265808	-30.48	0.000	-.8623183	-.7581237
en_couple	-.1252294	.019664	-6.37	0.000	-.1637702	-.0866885
ddipl						
3	-.0674434	.035364	-1.91	0.057	-.1367555	.0018688
4	-.2627095	.0324658	-8.09	0.000	-.3263413	-.1990778
5	-.3680032	.0311346	-11.82	0.000	-.429026	-.3069804
6	-.4981455	.0381265	-13.07	0.000	-.5728721	-.4234189
7	-.8048283	.0309099	-26.04	0.000	-.8654106	-.744246
_cons	-1.957251	.1033402	-18.94	0.000	-2.159794	-1.754708

Résultats : effets marginaux du probit

```
Average marginal effects                        Number of obs     =     29,248
Model VCE     : OIM

Expression    : Pr(actif), predict()
dy/dx w.r.t.  : age NBENF3 en_couple 3.ddipl 4.ddipl 5.ddipl 6.ddipl 7.ddipl
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0128503	.0001954	-65.76	0.000	-.0132333	-.0124673
NBENF3	-.2043994	.0064625	-31.63	0.000	-.2170657	-.1917331
en_couple	-.0315924	.0049528	-6.38	0.000	-.0412996	-.0218851
ddipl						
3	-.0146067	.0076632	-1.91	0.057	-.0296264	.000413
4	-.0609368	.0074383	-8.19	0.000	-.0755155	-.046358
5	-.0883387	.0072784	-12.14	0.000	-.1026041	-.0740732
6	-.124402	.0098168	-12.67	0.000	-.1436425	-.1051614
7	-.217612	.0079918	-27.23	0.000	-.2332756	-.2019484

Note: dy/dx for factor levels is the discrete change from the base level.

- Accepte-t-on l'hypothèse nulle du modèle sans explicative ?

  Réponse : On rejette $H_0 : \beta_{0-1} = 0$ en utilisant le test de rapport de vraisemblance. Pour le modèle Logit (page 40), la valeur de la statistique ξ_n^R vaut 8005, le p-value vaut 0.000, on rejette fortement H_0. Idem pour le modèle Probit.
- Quelles sont les variables significatives à 5% ?

  Réponse : Toutes les variables sont significatives à 5 % dans le cas du modèle Logit. Dans le cas du Probit seule la variable indiquant que DDIPL=3 ne l'est pas.
- Quelles sont les formules utilisées pour calculer les effet marginaux moyens et à la moyenne de la variable âge ?

  Réponse : On calcule l'effet marginal moyen de la variable âge via la formule

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i'\widehat{\beta})\left(\widehat{\beta}_{age} + 2\widehat{\beta}_{agecarre}\ age_i\right).$$

  Pour calculer l'effet marginal à la moyenne, on utilise la même formule sauf que les variables x_i et age_i sont remplacées par la moyenne de ces variables dans l'échantillon.

Questions sur l'application

- Estime-t-on ici l'effet causal des variables ?

  Réponse : On estime l'effet causal des variables si le modèle est bien spécifié et si ϵ et X sont bien indépendantes.

- Comment calcule-t-on le rapport des risques (odds-ratio) correspondant à une variable continue ?

  Réponse : Comme dans le cas d'une variable explicative binaire (voir page 14). Par exemple, pour la variable âge, le odds-ratio dans le tableau à la page 45 correspond à $e^{\widehat{\beta}_{age}}$ où $\widehat{\beta}_{age}$ est l'estimation du paramètre associé à l'âge.

- Les résultats des deux modèles sont-ils comparables ?

  Réponse : Oui, toutes les variables sont significatives dans les deux modèles (sauf celle correspondant à DDIPL==3), et les coefficients estimés ont le même signe. Les effets marginaux sont par ailleurs très proches.

L'essentiel

- Modélisation via les variables latentes. Modèles logit, probit.
- Effets marginaux : dérivation, interprétation.
- Odds-ratios pour le logit.
- Identification (normalisation), estimation via le max. de vrais.
- Choix de modèle via les critères d'information.
- Modèle de probabilité linéaire : intérêt, limites.