Correction – Quiz 3 sur le Chapitre 3
(modèles binaires)

(L.G.) – Cette version : 18 mars 2022

ENSAE 2A – Économétrie 2 – Printemps 2022

Sauf mention contraire, les notations utilisées cherchent à reprendre celles du cours.

Question 1 (effets marginaux, modèles linéaires ou binaires)

$X = (1, X_1, \dots, X_{K-1})'$ désignera le vecteur aléatoire des covariables = régresseurs = variables explicatives. On notera $\beta_0 = (\beta_{00}, \beta_{01}, \dots, \beta_{0K-1})'$ le vecteur des coefficients associés.

On considère ici une variable explicative d'intérêt continue X_1 et on s'intéresse, pour différents modèles, à l'effet de X_1 sur Y. X_2 est une autre variable explicative.

Y est la variable aléatoire réelle de résultat = variable expliquée. Elle pourra être continue ou binaire = dichotomique selon les modèles.

Dans chacun des modèles suivants, de (a) à (f),

1. Calculez *l'effet marginal de X_1 (sur Y)*[1] c'est-à-dire, la dérivée de l'application partielle qui à x_1 associe $\mathbb{E}[Y \mid X = x]$, les autres variables explicatives éventuelles étant fixées. On note parfois X_{-1} pour désigner le vecteur de ces variables explicatives en excluant le régresseur X_1 considéré : $x_1 \mapsto \mathbb{E}[Y \mid X = x] = \mathbb{E}[Y \mid X_1 = x_1, X_{-1} = x_{-1}]$.
2. Est-ce que cet effet marginal dépend de la valeur x des régresseurs ?

   Cadre général (modèles linéaires ou binaires) : on a des observations supposées i.i.d. (Y, X), ayant une certaine loi jointe. On s'intéresse à l'effet de X sur Y (pour un objectif de prédiction ou pour estimer un effet causal), et on est ainsi intéressé par apprendre la loi conditionnelle de Y sachant X (ou du moins certaines caractéristiques de cette loi conditionnelle). On pourrait essayer d'estimer directement cette loi mais c'est difficile car il s'agit en général d'un objet infini-dimensionnel, requérant a priori une estimation non paramétrique. Une autre approche est donc de restreindre les possibilités par un modèle en spécifiant une certaine relation entre Y et X via l'espérance conditionnelle de Y à X. Ainsi, un modèle binaire tel que vu dans le Chapitre 3 du cours revient à supposer

$$\mathbb{E}[Y \mid X] \overset{\text{car } Y \text{ binaire}}{=} \mathbb{P}(Y = 1 \mid X) \overset{\text{hypothèse de modélisation}}{=} F(X'\beta_0)$$

   pour une certaine fonction F *connue*, de sorte que le modèle est paramétrique : le vecteur β_0 est fini-dimensionnel (de dimension le nombre de composantes du vecteur X des régresseurs). Rappel : puisque $Y \in \{0, 1\}$, $Y = \mathbb{1}\{Y = 1\}$ et on a par conséquent $\mathbb{E}[Y \mid X] = \mathbb{E}[\mathbb{1}\{Y = 1\} \mid X] = \mathbb{P}(Y = 1 \mid X)$.

1. On précise parfois "l'effet marginal de X_1 **sur Y**", et parfois non, en laissant alors implicite le fait qu'il s'agit de l'effet sur la variable expliquée étudiée Y. *Pour les modèles non-linéaires, il pourra toutefois être intéressant de bien préciser quelle est la variable expliquée considérée* (on pourra en effet, parfois, distinguer l'effet sur Y, la variable observée, et l'effet sur Y^*, la variable latente).

Si F est l'identité, on suppose un modèle linéaire sur Y ("modèle linéaire" au sens habituel avec l'hypothèse que le résidus est indépendant en moyenne/espérance ("mean-independent") des régresseurs, modèle appelé "modèle de probabilité linéaire" lorsque $Y \in \{0, 1\}$:

$$\mathbb{E}[Y \mid X] = X'\beta_0.$$

Cette dernière équation est équivalente à écrire qu'il existe une variable aléatoire ε telle que

$$Y = X'\beta_0 + \varepsilon \quad \text{et} \quad \mathbb{E}[\varepsilon \mid X] = 0 \text{ (on dit que } \varepsilon \text{ est "mean-independent" de } X).$$

Dans tous les cas (qu'importe F et que Y soit continue ou binaire), ***l'effet marginal*** d'une composante quelconque[2] *continue* X_j de X **sur** Y **est défini comme la dérivée partielle par rapport à** x_j **de l'espérance conditionnelle de** Y **sachant** $X = x$ (on décompose x entre la j-ème composante x_j et les autres, notées x_{-j}).

Remarques et explications sur cette définition :

— On considère une **dérivée** car on s'intéresse à un effet *marginal*. Pour un régresseur X_j binaire (idem pour un X_j discret/catégoriel plus largement), la notion d'un changement marginal de X_j n'a pas le même sens puisqu'il n'y a que deux valeurs possibles. L'effet marginal est donc remplacé par *la différence discrète* suivante :

$$\mathbb{E}[Y \mid X_{-j} = x_{-j}, X_j = 1] - \mathbb{E}[Y \mid X_{-j} = x_{-j}, X_j = 0]$$

— Le fait de considérer une dérivée **partielle** renvoie à l'idée d'un effet de X_j sur Y *"toutes choses égales par ailleurs"* : on regarde l'effet d'une variation de X_j sur Y en gardant les autres composantes de X fixées[3] $\longrightarrow$ a priori, la valeur de cet effet marginal dépend donc évidemment de la valeur x_{-j} fixée des autres composantes puisqu'on considère la dérivée d'une application partielle.

— De surcroît, **la dérivée est une notion locale** $\longrightarrow$ a priori, la valeur de l'effet marginal dépend donc également de la valeur x_j à laquelle est évaluée la dérivée partielle.

L'effet marginal est donc en général **une fonction** qui dépend de x_j et de x_{-j}, donc de x.

L'exception à la règle est le cas d'un modèle linéaire "simple"[4] où il s'avère que l'effet marginal ne dépend ni de la valeur de x_j, ni de la valeur x_{-j} des autres régresseurs.

Il faut vraiment comprendre qu'il s'agit d'un cas très spécial et qu'en général, et donc en particulier pour les modèles binaires et pour des modèles linéaires plus généraux / sophistiqués, l'effet marginal dépend de la valeur x des régresseurs.

3. Donnez l'expression de *l'effet marginal moyen de* X_1 *(sur* Y*)*, c'est-à-dire l'espérance de l'effet marginal de X_1 sur Y, où l'espérance porte sur les régresseurs X. On peut l'écrire de manière générale ainsi :

$$\text{effet marginal moyen de } X_1 \text{ sur } Y \;=\; \mathbb{E}_X\left[\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1}\right],$$

où l'indice X sous l'espérance $\mathbb{E}_X$ est employée pour expliciter le fait que l'aléa porte ici sur la variable X.

2. On considère dans les exemples qui suivent la première composante X_1.

3. On veut aussi garder le terme d'erreur ε fixé, d'où les problématiques d'endogénéité et d'exogénéité et du lien entre X et ε ; en lien avec la question effet causal ou simple corrélation.

4. Où " simple" signifie ici sans interactions entre les variables explicatives ni de puissances ou d'autres transformations, les variables explicatives apparaissent seulement en niveau, à la puissance 1 – voir le modèle (a) ci-dessous.

(a) Modèle linéaire (cas simple) Y continue, $X = (1, X_1, X_2)'$ et

$$\mathbb{E}[Y \mid X] = X'\beta_0 = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2,$$

ce qu'on peut écrire de façon équivalente [5]

$$Y = X'\beta_0 + \varepsilon \text{ avec } \mathbb{E}[\varepsilon \mid X] = 0.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = \beta_{01}.$$

L'effet marginal, dans ce cas très particulier, ne dépend pas de x ; c'est une fonction constante égale à β_{01}. Par conséquent, l'effet marginal moyen est simplement égal à

$$\mathbb{E}_X\left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1}\right) = \mathbb{E}_X[\beta_{01}] = \beta_{01}.$$

De même, l'effet marginal à la moyenne, ou d'ailleurs l'effet marginal en tout autre point x, est également égal à β_{01}, de même que l'effet marginal moyen pour tout sous groupe.

(b) Modèle linéaire (avec des puissances, par exemple un effet quadratique de X_1) Y continue, $X = (1, X_1, X_1^2, X_2)'$ et

$$\mathbb{E}[Y \mid X] = X'\beta_0 = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_1^2 + \beta_{03}X_2.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = \beta_{01} + 2x_1\beta_{02}.$$

L'effet marginal dépend donc de x. Plus précisément, dans un modèle linéaire sans interaction entre différents régresseurs (uniquement des puissances de X_1), l'effet marginal de X_1 ne dépend pas de la valeur x_{-1} des autres régresseurs mais est bien une fonction de x_1. Cela permet de prendre en compte un effet quadratique et non simplement linéaire de X_1 sur Y.

L'effet marginal moyen est alors

$$\mathbb{E}_X\left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1}\right) = \mathbb{E}_X[\beta_{01} + 2X_1\beta_{02}] = \beta_{01} + 2\beta_{02}\mathbb{E}[X_1]$$

où la dernière égalité utilise la linéarité de l'espérance. L'effet marginal moyen dépend donc des caractéristiques de la population considérée, ici de $\mathbb{E}[X_1]$. Dans un tel modèle linéaire, remarquons que l'effet marginal moyen est aussi égal à l'effet marginal à la moyenne.

5. *Il est important de bien comprendre que ces deux écritures sont équivalentes ; elles postulent un modèle sur l'espérance conditionnelle de Y sachant X.* On ne réécrira pas systématiquement cette équivalence dans les modèles linéaires qui suivent, on écrira seulement une des deux formulations équivalentes.

(c) Modèle linéaire (avec des interactions – ici, un terme dit d'interaction sans "main effect" de X_1) Y continue, $X = (1, X_1 \times X_2, X_2)'$ et

$$Y = \beta_{00} + \beta_{01} X_1 X_2 + \beta_{02} X_2 + \varepsilon \text{ avec } \mathbb{E}[\varepsilon \mid X] = 0.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = \beta_{01} x_2.$$

L'effet marginal dépend donc de x et il dépend en fait de la valeur $x_{-1} = x_2$ ici des autres régresseurs seulement, et non de x_1. L'effet marginal moyen est alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} x_2] = \beta_{01} \mathbb{E}[X_2]$$

par linéarité de l'espérance. Il est aussi égal à l'effet marginal à la moyenne ici comme dans le modèle (c).

(d) Modèle linéaire (avec des interactions – ici, un terme dit d'interaction et un "main effect" de X_1) Y continue, $X = (1, X_1, X_2, X_1 \times X_2)'$ et

$$Y = \beta_{00} + \beta_{01} X_1 + \beta_{02} X_2 + \beta_{03} X_1 X_2 + \varepsilon \text{ avec } \mathbb{E}[\varepsilon \mid X] = 0.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = \beta_{01} + \beta_{03} x_2.$$

L'effet marginal dépend de x, ici encore, comme dans le modèle (c), uniquement via la valeur des autres régresseurs $x_{-1} = x_2$. L'effet marginal moyen est alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} + \beta_{03} x_2] = \beta_{01} + \beta_{03} \mathbb{E}[X_2]$$

par linéarité de l'espérance. Il est à nouveau égal à l'effet marginal à la moyenne.

(e) Modèle linéaire (un autre exemple) Y continue, $X = (1, X_1, X_2, X_1^2, X_1^2 \times X_2)'$ et

$$\mathbb{E}[Y \mid X] = \beta_{00} + \beta_{01} X_1 + \beta_{02} X_2 + \beta_{03} X_1^2 + \beta_{04} X_1^2 X_2.$$

Qu'importe la complexité du modèle (en pratique, il faut réfléchir et se demander quelle forme fonctionnelle pour $\mathbb{E}[Y \mid X]$ est la plus pertinente pour chaque situation, selon ce que sont Y et X), il suffit pour le calcul de suivre les définitions. Ainsi, pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a ici

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = \beta_{01} + 2\beta_{03} x_1 + 2\beta_{04} x_1 x_2$$

L'effet marginal dépend de x, à la fois de x_1 et de $x_{-1} = x_2$. L'effet marginal moyen est alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} + 2\beta_{03} x_1 + 2\beta_{04} x_1 x_2] = \beta_{01} + 2\beta_{03} \mathbb{E}[X_1] + 2\beta_{04} \mathbb{E}[X_1 X_2]$$

par linéarité de l'espérance. Il dépend ici d'un moment croisé, $\mathbb{E}[X_1X_2]$, donc de la covariance entre X_1 et X_2. En général, $\mathbb{C}\text{ov}(X_1, X_2) \neq 0$, c'est-à-dire, $\mathbb{E}[X_1X_2] \neq \mathbb{E}[X_1]\mathbb{E}[X_2]$. Par conséquent ici, même s'il s'agit d'un modèle linéaire, l'effet marginal moyen n'est pas égal en général à *l'effet marginal à la moyenne*, définie comme l'effet marginal évalué en $x = \mathbb{E}[X] = (1, \mathbb{E}[X_1], \mathbb{E}[X_2])'$ ici,[6] et qui vaut donc ici

$$\text{Effet marginal à la moyenne : } \left.\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_j}\right|_{x=\mathbb{E}[X]} = \beta_{01} + 2\beta_{03}\mathbb{E}[X_1] + 2\beta_{04}\mathbb{E}[X_1]\mathbb{E}[X_2]$$
$$\neq \beta_{01} + 2\beta_{03}\mathbb{E}[X_1] + 2\beta_{04}\mathbb{E}[X_1X_2]$$

en général, dès lors que X_1 et X_2 sont corrélés.

(f) Modèle binaire (cas simple) Y binaire, $X = (1, X_1, X_2)'$ et

$$\mathbb{E}[Y \mid X] = \mathbb{P}(Y = 1 \mid X) = F(X'\beta_0) = F(\beta_{00} + \beta_{01}X_1 + \beta_{02}X_2),$$

où F est une fonction de répartition connue (Φ par exemple pour un modèle probit ou Λ pour un modèle logit).

De façon équivalente (c'est le message de la slide 7 du Chapitre 3), on peut formuler cette hypothèse sur l'espérance conditionnelle de Y sachant X, qui est ici, puisque Y est binaire ($Y = \mathbb{1}\{Y = 1\}$), égale à la probabilité conditionnelle de $Y = 1$ sachant X au moyen d'une variable dite *latente*, notée Y^* :

$$Y = \mathbb{1}\{Y^* \geq 0\} \text{ et } Y^* = X'\beta_0 + \varepsilon = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \varepsilon.$$

avec $-\varepsilon$ une variable aléatoire réelle indépendante de X et dont la distribution a pour fonction de répartition F.

On suit la même méthode et les mêmes définitions ; il faut juste faire **attention** puisqu'on a la **dérivée d'une fonction composée**. Comme dans le cours, notons f la dérivée de F. Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} = f(\beta_{00} + \beta_{01}x_1 + \beta_{02}x_2) \times \beta_{01} = f(x'\beta_0)\beta_{01}.$$

Ainsi, à la différence des modèles linéaires, même dans un cas "simple" (pas d'interaction, de puissance ou autre transformation des variables explicatives), l'effet marginal dépend de x, à la fois de x_1 et de la valeur x_{-1} des autres régresseurs car x intervient dans f du fait de la dérivée d'une fonction composée. Ce sera toujours le cas dans les modèles binaires et plus largement pour les modèles non-linéaires vus dans la suite du cours d'Économétrie 2.

L'effet marginal moyen est égal à

$$\mathbb{E}_X\left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1}\right) = \mathbb{E}_X\left[f(X'\beta_0)\beta_{01}\right] = \mathbb{E}[f(X'\beta_0)]\beta_{01}$$

par linéarité de l'espérance. Il est différent de l'effet marginal à la moyenne, égal à

$$f(\mathbb{E}[X]'\beta_0)\beta_{01}$$

On retrouve dans ce cas simple les formules du cours (Chapitre 3, slides 12 et 13).

6. Rappel : l'espérance s'applique composante par composante

(f bis) *Questions supplémentaires pour le modèle binaire (f).*

— Quel est l'effet marginal de X_1 *sur* Y^* (et non sur Y, attention !) dans le modèle (f) ?

Dans un modèle binaire, le modèle concernant la variable latente Y^* est un modèle linéaire. La définition d'un effet marginal reste le même sauf qu'on regarde maintenant Y^* au lieu de Y. On a ainsi, pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$,

$$\frac{\partial \mathbb{E}[Y^* \,|\, X = x]}{\partial x_1} = \beta_{01}.$$

La question est pour simplement pour rappeler que considérer Y ou Y^* comme variable d'intérêt sont deux choses différentes, deux variables d'intérêt distinctes. ***Dans un modèle binaire****, on observe Y et non Y^* donc l'**effet marginal qu'on considère naturellement est l'effet marginal sur** Y, la variable binaire **observée**.*

Remarque importante : parfois il n'y a pas de sens à regarder Y^*, on ne pourra jamais l'observer ou bien la variable n'a pas de sens quantitatif précis – appelons cela le cas (i). Parfois, au contraire, cela pourrait faire sens de regarder l'effet de X_1 sur Y^* : Y^* a un sens quantitatif clair et on aurait aimé observer Y^* et estimer une régression de Y^* sur X mais, pour des raisons de limitation de données, on a seulement accès à Y et on doit se contenter d'un modèle binaire – notons ce cas le cas (ii). [7]

— Pour les quatre exemples étudiés aux slides 8 et 9 du Chapitre 3, discuter s'il s'agit plutôt de la situation (i) ou (ii) pour la variable latente Y^*.

Les exemples 2 et 4 appartiennent au cas (ii). Dans ces cas, on pourrait avoir des mesures du niveau de dette nette de chaque entreprise ou la moyenne générale de chaque élève. La variable latente a un sens quantitatif clair et bien défini et on pourrait l'observer si on la mesurait. Remarque : malgré cela, même si on observait Y^*, on pourrait vouloir s'intéresser aussi à Y.

L'exemple 1 est clairement dans le cas (i) : on ne pourra jamais observer l'utilité pour chaque individu, c'est une valeur seulement ordinale et non cardinale. Une façon de voir cela en cas de doute : se demander quelle est l'unité, la dimension (au sens d'unité physique) de Y^* ? S'il n'y a pas de réponse claire, d'unité courante, c'est en général que Y^* est une véritable variable latente, au sens où on ne saurait l'observer et elle intervient uniquement comme une façon alternative de modéliser l'hypothèse sur $\mathbb{P}(Y = 1 \,|\, X)$. Ici, quelle unité pour l'utilité ? Aucune idée ! L'exemple 3 est un peu intermédiaire au sens où cela paraît impossible techniquement de dénombrer et vraiment de compter ainsi le nombre de bactéries mais pourquoi pas en théorie. [8]

Au delà de ces exemples particuliers, l'idée est juste de comprendre que

1. les effets marginaux sur Y ou sur Y^* sont deux choses différentes ;
2. **en règle générale, la variable d'intérêt dans un modèle binaire est la variable observée binaire Y**
3. dans certains cas néanmoins (si elle a un sens quantitatif naturel), on pourrait vouloir mesurer et s'intéresser également à la variable sous-jacente Y^* (mais on ne l'observe pas dans les modèles binaires).

7. Voir également la Question 2 (a) de ce Quiz.

8. En fait, cette remarque pourrait même s'appliquer, en poussant un peu, à l'exemple 1. L'évolution technologique, scientifique est impossible à prévoir et il se pourrait que dans quelques années, la neuro-économie et les progrès dans la compréhension du cerveau amènent à être capable de mesurer véritablement une utilité des agents, qui se compterait, mettons, en intensité des liaisons électriques entre neurones, au nombre de synapses connectées ou quelque chose d'autre dont on n'a aucune idée actuellement.

(g) Modèle binaire (un autre exemple plus compliqué avec un "main effect", un terme d'interaction et un effet quadratique) Y binaire, $X = (1, X_1, X_2, X_1^2, X_1X_2)'$ et

$$\mathbb{E}[Y \mid X] = \mathbb{P}(Y = 1 \mid X) = F(X'\beta_0) = F(\beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_1^2 + \beta_{04}X_1X_2).$$

On suit toujours la même méthode, il faut juste être attentif au calcul de la dérivée. Ici, pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, l'effet marginal de X_1 sur Y est égal à

$$\begin{aligned}\frac{\partial \mathbb{E}[Y \mid X = x]}{\partial x_1} &= f(\beta_{00} + \beta_{01}x_1 + \beta_{02}x_2 + \beta_{03}x_1^2 + \beta_{04}x_1x_2) \times (\beta_{01} + 2\beta_{03}x_1 + \beta_{04}x_2) \\ &= f(x'\beta_0) \times (\beta_{01} + 2\beta_{03}x_1 + \beta_{04}x_2)\,.\end{aligned}$$

Il dépend donc de x. L'effet marginal moyen vaut

$$\mathbb{E}_X\left(\frac{\partial \mathbb{E}[Y \mid X]}{\partial X_1}\right) = \mathbb{E}_X\left[f(X'\beta_0)\,(\beta_{01} + 2\beta_{03}X_1 + \beta_{04}X_2)\,\right].$$

Question 2 (identification)

On considère le modèle binaire du Chapitre 3

$$\mathbb{E}[Y \mid X] = \mathbb{P}(Y = 1 \mid X) = F(X'\beta_0),$$

où F est une fonction de répartition connue. De façon équivalente,

$$Y = \mathbb{1}\{Y^* \geq 0\} \text{ et } Y^* = X'\beta_0 + \varepsilon$$

avec $-\varepsilon$ une variable aléatoire réelle indépendante de X et dont la distribution a pour fonction de répartition F.

Le paramètre d'intérêt qu'on cherche à identifier et estimer est β_0. On observe pour cela un échantillon i.i.d $(Y_i, X_i)_{i=1,\dots,n} \sim (Y, X)$.

(a) Si Y^* était observée Imaginons qu'on observe en fait également la variable latente, c'est-à-dire qu'on dispose d'un échantillon i.i.d. $(Y_i^*, Y_i, X_i)_{i=1,\dots,n} \sim (Y^*, Y, X)$.

Sous quelles conditions et comment peut-on identifier et estimer de façon consistante β_0 dans ce cas ?

L'idée : si on observe Y^*, il suffit de faire la régression linéaire de Y^* sur X. Celle-ci va bien identifier β_0 si et seulement si l'écriture $Y^* = X'\beta_0 + \varepsilon$ est la représentation "projection linéaire", c'est-à-dire, s'il n'y a pas de problème d'endogénéité de X, c'est-à-dire, si on a bien $\mathbb{E}[X\varepsilon] = 0$. [9]

Ici, on a supposé que $-\varepsilon$, et donc également ε, est indépendant de X. ε est donc a fortiori non corrélé avec toutes les composantes de X. Il suffit donc de supposer que ε est centré (ce qui est le cas par exemple dans un probit ou un logit et est le cas sans perte de généralité dès lors que X admet une constante) ainsi que les conditions de moments habituelles de la projection linéaire :

1. $\mathbb{E}[|Y^*|^2] < +\infty$ (*moment d'ordre 2 fini pour la variable expliquée*) ;

9. Rappel : X est un vecteur colonne, une variable aléatoire réelle donc l'égalité précédente signifie que pour toute composante X_j de X, on a $\mathbb{E}[X_j\varepsilon] = 0$. Si X inclut une constante, $\mathbb{E}[X\varepsilon] = 0$ est donc équivalent à ε est centré et non corrélé avec toutes les composantes de X.

2. $\mathbb{E}[\|X\|^2] < +\infty$ où $\|X\|$ est la norme euclidienne de X, de façon équivalente, pour toute composante X_j de X, $\mathbb{E}[|X_j|^2] < +\infty$ (*moment d'ordre 2 fini pour les variables explicatives*) ;
3. $\mathbb{E}[XX']$ est inversible (*pas de multicolinéarité parfaite des variables explicatives*).

Si les conditions de moments 1, 2 et 3 sont vérifiées et si ε est centré, β_0 est identifié par la régression linéaire de Y^* sur X et l'estimateur MCO de cette régression, qui est égal à (**Rappel : formule du cours d'Économétrie 2, Chapitre 1, slide 5** – ici avec Y^* comme variable expliquée)

$$\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i^*\right) = \left(\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sum_{i=1}^n X_i Y_i^*\right),$$

est un estimateur consistant de β_0.

(b) Relisez votre cours de Statistique 1 et rappelez la définition d'un modèle statistique identifiable.

Rappel : l'identifiabilité d'un modèle statistique a trait à l'injectivité de la fonction qui à une valeur possible du paramètre du modèle associe la loi des observations. On dit que le modèle est identifiable si cette fonction est injective. Autrement dit, si, pour une distribution quelconque des observations, cette distribution correspond au plus à un seul paramètre, le modèle est identifiable.

Tout le point de la slide 16 du Chapitre 3 est de montrer qu'on obtient la même distribution pour les observations (Y, X) avec différents paramètres : une distribution des observations correspond à plusieurs valeurs possibles du paramètre. La fonction n'est donc pas injective ; la distribution a plusieurs antécédents par cette fonction.

Remarque : on considère dans le Chapitre 3 *le modèle conditionnel aux régresseurs*. Dans ce modèle conditionnel aux X, on considère $X_1, \ldots, X_n$ fixés et tout se passe comme si on écrivait le modèle statistique (et ensuite la vraisemblance pour calculer l'Estimateur du Maximum de Vraisemblance) pour les observations $Y_1, \ldots, Y_n$ seulement.

Implicitement, on fait l'hypothèse que β_0 ne dépend pas de la loi des régresseurs X, ne dépend que de la distribution de Y conditionnelle à X et c'est pourquoi on peut considérer le modèle conditionnel aux régresseurs.

Le résultat d'identification du Chapitre 3 est énoncé dans le Théorème 1 (slide 17).

Si le seuil s et l'écart-type σ_0 du résidus sont fixés et si $\mathbb{E}[XX']$ est inversible, alors le modèle est identifié (toute distribution des observations est associé à un unique paramètre β_0).

De façon implicite mais à néanmoins bien garder en tête, ce résultat suppose également que le modèle est bien spécifié [10], c'est-à-dire qu'on a bien

$$\mathbb{E}[Y \mid X] = F(X'\beta_0)$$

pour une fonction de répartition F connue, ou, de façon équivalente, avec la modélisation par une variable latente :

$$Y = \mathbb{1}\{Y^* \geq 0\} \text{ et } Y^* = X'\beta_0 + \varepsilon$$

avec $-\varepsilon$ une variable aléatoire réelle indépendante de X et dont la distribution a pour fonction de répartition F. Dit autrement, on ne sait pas tromper sur le choix de la fonction F : les données sont véritablement générées selon ce modèle utilisant la fonction F supposée être connue.

10. Voir également TD 5, Exercice 2 pour ces questions de bonne ou mauvaise spécification du modèle.

(c) Rappelez le sens de l'hypothèse "$\mathbb{E}[XX']$ inversible". Aurait-on également besoin de cette hypothèse si l'on observait Y^* (question (a)) ?

Cette hypothèse signifie que les variables explicatives ne sont pas parfaitement corrélées (corrélation égale à 1 ou à -1). Rappel : cela n'interdit pas que les différentes composantes de X soient corrélées mais la corrélation ne peut être parfaite parfaite. [11]

On aurait également besoin de cette hypothèse pour identifier β_0 si on observait Y^*. Par contre dans ce cas (voir question (a)), on n'a pas besoin de supposer connus le seuil s et la variance du résidus ε du modèle sur la variable latente. De façon plus fondamentale, si l'on observait Y^*, on n'aurait pas besoin de supposer une loi connue pour ε ni l'indépendance entre ε et les régresseurs X. [12]

Il faut retenir de ce résultat d'identification qu'on ne peut pas identifier β_0 sans supposer le seuil s et la variance σ^2 du résidus ε connus. Dit autrement, on ne peut pas connaître/identifier avec l'observation de la loi de (Y, X) le paramètre β_0 mais seulement β_0/σ : β_0 à un coefficient proportionnel (positif) près. [13]

Intuitivement, cela explique les trois grands messages sur l'interprétation d'un coefficient β_{0j} de β_0 (slides 11 et 12 du Chapitre 3) – voir Question 7 de ce Quiz.

(d) Identification de s et σ^2 si Y^* était observée Imaginons comme en question (a) qu'on observe la variable latente Y^*. On suppose le modèle suivant, pour $s \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}_+^*$:

$$Y = \mathbb{1}\{Y^* \geq s\} \text{ et } Y^* = X'\beta_0 + \varepsilon$$

avec $-\varepsilon$ une variable aléatoire réelle indépendante de X et dont la distribution a pour fonction de répartition F et $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{V}[\varepsilon] = \mathbb{E}[\varepsilon^2] = \sigma^2$.

Avec un échantillon i.i.d. $(Y_i^*, Y_i, X_i)_{i=1,\dots,n} \sim (Y^*, Y, X)$,

— Est-il possible d'identifier et d'estimer de façon consistante s et σ^2 ?

— Si oui, comment ?

— Mêmes questions si l'on observait seulement X et Y^*, mais pas Y.

Il faut remarquer qu'on n'a pas besoin de supposer F connue dans ce cas si Y^* était observé. Il suffit de supposer les trois conditions de moments de la question (a) et l'exogénéité de X dans le modèle linéaire de la variable latente $Y^* = X'\beta_0 + \varepsilon$, c'est-à-dire, $\mathbb{E}[X\varepsilon] = 0$.

Supposons ici, comme habituellement en l'absence de précision contraire, que le vecteur X contient une constante. L'hypothèse $\mathbb{E}[\varepsilon] = 0$ centré est alors sans perte de généralité. ε étant centré, $\mathbb{E}[X\varepsilon] = 0$ est équivalent à l'absence de corrélation entre ε et toute composante X_j de X. On a bien cela par hypothèse puisqu'on suppose a fortiori $\varepsilon \perp\!\!\!\perp X$.

Ainsi, comme en (a), la régression linéaire de Y^* sur X identifie β_0 et l'estimateur MCO $\widehat{\beta}$ de Y^* sur X est un estimateur consistant de β_0.

Avec $\widehat{\beta}$, on peut récupérer dans un second temps les résidus estimés : $\widehat{\varepsilon}_i := Y_i^* - X_i'\widehat{\beta}$ pour toute observation $i \in \{1, \dots, n\}$.

On a supposé $\varepsilon \perp\!\!\!\perp X$. A fortiori, la variance conditionnelle à X de ε ne dépend donc pas de X : le résidus ε est ici homoscédastique.

11. Pour plus de détails sur cette condition, revoir le Chapitre 1 d'Économétrie 1 (slide 18 notamment).

12. Sur ce dernier point et la différence entre modèle paramétrique ou semi-paramétrique, voir également la question 8 sur les modèles.

13. Pour le coefficient de la constante β_{00}, c'est encore pire au sens où on ne le connaît de surcroît que modulo une translation s ; en gros, on ne le connaît pas du tout donc, mais ce n'est pas grave dans la mesure où l'on ne s'intéresse généralement pas à la constante.

On peut alors estimer la variance de ε (= son moment d'ordre deux car centré) de façon classique (voir par exemple EM1, TD1)) par la moyenne empirique des résidus estimés au carré :

$$\widehat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

ou, pour avoir un estimateur sans biais à distance finie, avec K la dimension du vecteur X,

$$\widehat{\sigma}_n^2 := \frac{1}{n-K}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

On pourrait donc identifier et estimer σ^2 de façon consistante si Y^* était observé.

En pratique bien sûr, dans des modèles binaires, on n'observe pas Y^* donc on ne peut pas faire cela. La question était pour faire quelques rappels et pour insister sur le fait que la nécessité des hypothèses plus fortes pour identifier β_0 dans un modèle binaire, par opposition à un modèle linéaire, vient de cette perte d'information : on observe seulement Y binaire, c'est-à-dire si Y^* continu est au-dessus ou en dessous d'un certain seuil et non les valeurs de Y^*.

Remarque : pour l'estimation de σ^2, on n'a pas eu besoin de Y, on pourrait donc faire de exactement pareil en observant seulement X et Y^*.

[*Point plus secondaire*] Regardons maintenant le paramètre de seuil s. Dans un modèle binaire (sans observer Y^* observé), on doit le supposer connu et il est habituellement fixé à 0.

Ici, c'est un problème d'estimation pas tout à fait standard mais d'un genre similaire à celui que vous avez rencontré lors de l'estimation du support d'une loi uniforme (voir exercice de Statistiques 1 sur le modèle Uniforme($[0,\theta]$)).

Par hypothèse, on sait que $Y = \mathbb{1}\{Y^* \geq s\}$ et on observe un échantillon i.i.d. des variables Y et Y^*. L'idée est alors simplement de prendre, par exemple, comme estimateur de s la plus petite valeur de Y^* tel que $Y = 1$:[14]

$$\widehat{s}_{(m)} := \min\{Y_i^*, i \in \{1,\dots,n\} : Y_i = 1\}$$

On peut aussi prendre la plus grande valeur observée de Y^* dans les données avant de dépasser le seuil :

$$\widehat{s}_{(M)} := \max\{Y_i^*, i \in \{1,\dots,n\} : Y_i = 0\}$$

Une autre possibilité permettant d'utiliser ces deux informations est de prendre la moyenne de ces deux estimateurs et de poser

$$\widehat{s} := \frac{\widehat{s}_{(m)} + \widehat{s}_{(M)}}{2}.$$

Enfin, Y dépend de s : on ne peut plus faire cela et on n'a aucune idée du seuil s si l'on observe Y^* mais non Y, ce qui est logique.

Question 3 (identification et estimation)

On considère le modèle binaire suivant

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

avec[15] $\varepsilon \,|\, X \sim \mathcal{N}(0,\sigma^2)$, $s \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}_+^*$.

Alors

14. Voir le script Stata de correction du TD 5, Exercice 1 pour la mise en oeuvre de ces estimateurs et la vérification par simulation de leur consistance.

15. Attention : $\varepsilon \,|\, X \sim \mathcal{N}(0,\sigma^2)$ est une façon raccourcie d'écrire que (i) la loi conditionnelle de ε sachant X ne dépend pas de X, donc $\varepsilon \perp\!\!\!\perp X$, et (ii) que la loi de ε est une normale $\mathcal{N}(0,\sigma^2)$. Ici c'est une loi symétrique donc ε et $-\varepsilon$ ont même loi (pour faire le lien avec le cours où on spécifie, dans le cas général, la loi de $-\varepsilon$).

1. l'estimateur des moindres carrés ordinaires (MCO) de Y sur X est un estimateur consistant de β_0 – **Faux**, c'est justement tout le point du Chapitre 3 et des modèles binaires, il faut faire autre chose (un modèle probit ici vu les hypothèses) que des MCO pour estimer β_0.
2. l'estimateur des MCO de Y sur X est un estimateur consistant des effets marginaux moyens de X sur Y – **Faux**, voir Question 1.
3. l'estimateur du maximum de vraisemblance (EMV) de ce modèle est un estimateur consistant de β_0 à condition de normaliser s à 0 et σ^2 à 1 – **Vrai**, voir Question 2 et la discussion sur l'identification : il faut fixer le seuil et la variance des résidus pour identifier β_0. En fixant s à 0 et σ^2 à 1 (c'est-à-dire que $\varepsilon \,|\, X \sim \mathcal{N}(0,1)$) on est dans un modèle probit. L'estimateur probit de Y sur X est, par définition, l'estimateur du maximum de vraisemblance dans ce modèle (modèle conditionnel aux régresseurs) et, d'après le cours, c'est un estimateur consistant de β_0.
4. l'EMV de ce modèle est un estimateur consistant de β_0 même si le modèle est hétéroscédastique, c'est-à-dire si $\varepsilon \,|\, X \sim \mathcal{N}(0, \sigma^2(X))$ pour une fonction $\sigma^2(\cdot)$ quelconque – **Faux**, **dans les modèles binaires, on a besoin de supposer que le résidus ε du modèle linéaire sur la variable latente est indépendant de X**. Cette hypothèse implique en particulier que la variance de ε ne dépend pas de X. Autrement dit, elle implique l'homoscédasticité de ε. A nouveau, on voit que le fait de n'observer que $Y \in \{0,1\}$ et non Y^* continu oblige à des **hypothèses bien plus fortes sur le lien entre résidus ε et régresseurs X que dans un modèle linéaire** : il faut $\varepsilon \perp\!\!\!\perp X$ et donc a fortiori ε homoscédastique.

Question 4 (interprétation des paramètres)

Trois messages importants quant à l'interprétation du paramètre β_0 dans un modèle binaire : (slides 11, 12 et 13 du Chapitre 3)

1. **On peut interpréter *qualitativement*** (c'est-à-dire le signe de) β_{0j} : toutes choses égales par ailleurs, X_j a un effet positif (respectivement négatif) sur $\mathbb{P}(Y=1 \,|\, X)$ si et seulement si $\beta_{0j} > 0$ (resp. $\beta_{0j} < 0$). Explication de cela : pour X_j une composante "simple" (i.e., sans puissance ou interaction avec un autre régresseur), on a
$$\frac{\partial \mathbb{E}[Y \,|\, X = x]}{\partial x_j} = f(x'\beta_0)\beta_{0j}$$
Or, $f = F'$ est une densité donc à valeurs positives : pour tout x, le signe de $\frac{\partial \mathbb{E}[Y \,|\, X=x]}{\partial x_j}$, l'effet marginal de X_j sur Y, est égal au signe de β_{0j}, d'où l'interprétation qualitative valide sur le signe du coefficient.
2. Par contre, **on ne peut pas avoir d'interprétation *quantitative*** immédiate du coefficient β_{0j} quant à l'effet de X_j sur $\mathbb{P}(Y=1 \,|\, X)$ $\longrightarrow$ il faut passer par l'effet marginal de X_j : c'est une fonction de x et l'effet quantitatif dépend donc de x.
3. On peut par contre interpréter **quantitativement** le *ratio de deux coefficients* β_{0l}/β_{0j} – voir des exemples Question 7. L'idée est la suivante (voir dernier point du slide 12). Pour deux régresseurs X_j, X_l "simples" (c'est-à-dire pas de puissance ou d'interaction), pour tout x fixé, l'effet marginal de X_l sur Y divisé par l'effet marginal de X_j sur Y est égal à
$$\frac{\partial \mathbb{E}[Y \,|\, X = x]/\partial x_l}{\partial \mathbb{E}[Y \,|\, X = x]/\partial x_j} = \frac{f(x'\beta_0)\beta_{0l}}{f(x'\beta_0)\beta_{0j}} = \frac{\beta_{0l}}{\beta_{0j}}.$$

On considère le modèle binaire
$$Y = \mathbb{1}\{\beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \varepsilon \geq 0\}$$

avec ε indépendante de $X = (1, X_1, X_2)'$.

Alors [16]

1. β_{01} peut s'interpréter comme l'effet marginal moyen de X_1 sur Y – **Faux**, voir Question 1, (f) pour le calcul de l'effet marginal moyen dans ce modèle.
2. l'effet marginal moyen de X_2 sur Y est égal à l'effet marginal de X_2 sur Y à la moyenne – **Faux**, voir à nouveau Question 1, (f).
3. β_{02}/β_{01} peut s'interpréter comme le ratio de l'effet marginal moyen de X_2 sur l'effet marginal moyen de X_1 – **Vrai**, même calcul qu'au message 3. ci-dessus avec les effets marginaux moyens :

$$\frac{\mathbb{E}_X[\partial \mathbb{E}[Y \mid X]/\partial X_l]}{\mathbb{E}_X[\partial \mathbb{E}[Y \mid X]/\partial X_j]} = \frac{\mathbb{E}_X[f(X'\beta_0)\beta_{0l}]}{\mathbb{E}_X[f(X'\beta_0)\beta_{0j}]} = \frac{\mathbb{E}_X[f(X'\beta_0)]\,\beta_{0l}}{\mathbb{E}_X[f(X'\beta_0)]\,\beta_{0j}} = \frac{\beta_{0l}}{\beta_{0j}}.$$

4. les effets marginaux de X_1 et de X_2 ne sont définis que si ε suit une loi logistique ou une loi normale – **Faux**, absurdité, la notion d'effet marginal est défini de façon bien plus générale, y compris pour des modèles non binaires d'ailleurs (voir Question 1).
5. $\exp(\beta_{01})$ correspond au rapport des risques (odds-ratio) pour X_1 si ε suit une loi normale – **Faux**, les odds-ratio sont spécifiques au modèle logit, donc si ε suit une loi logistique et non normale.

Question 5 (estimation et optimisation)

On considère un modèle logit de Y sur $X = (1, X_1, X_2)'$ avec $X_1 \in \{0, 1\}$, c'est-à-dire

$$Y = \mathbb{1}\{\beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \varepsilon \geq 0\},$$

où ε suit une loi logistique ($F = \Lambda$).

Alors, la log-vraisemblance conditionnelle aux X de ce modèle

1. ne peut pas être maximisée si X_1 et X_2 sont corrélés entre eux – **Faux**, on peut avoir de la corrélation entre les régresseurs (il faut juste qu'elle ne soit pas parfaite).
2. ne peut pas être maximisée si X_1 ou X_2 sont corrélés avec ε, le résidus du modèle sous-jacent – **Faux**, on pourrait bien maximiser cette vraisemblance mais le modèle logit serait par contre mal-spécifié : il suppose $\varepsilon \perp\!\!\!\perp X$. Si ce n'est pas le cas, si ε et X sont corrélés, l'estimateur de l'EMV va bien converger en probabilité vers une certaine valeur mais cette limite en probabilité ne sera pas le paramètre d'intérêt $\beta_0 = (\beta_{00}, \beta_{01}, \beta_{02})'$. Dit autrement, dans ce cas, l'estimateur logit n'est pas un estimateur consistant de β_0 car le modèle est mal-spécifié.
3. peut être maximisée mais admet de multiples maxima si pour tout $i \in \{1, \ldots, n\}$, $X_{1i} = 1$ implique $Y_i = 1$ – **Faux**.
4. ne peut pas être maximisée si pour tout $i \in \{1, \ldots, n\}$, $X_{1i} = 1$ implique $Y_i = 1$, mais elle pourrait l'être si ε suivait une loi normale – **Faux**.
5. ne peut pas être maximisée si pour tout $i \in \{1, \ldots, n\}$, $X_{1i} = 1$ implique $Y_i = 1$, et elle ne pourrait pas l'être non plus quelle que soit la loi de ε (logistique, normale, ou autre) – **Vrai**, ces trois dernières propositions font directement référence au problème dit de "séparation complète" des données (slide 21, Chapitre 3). Dans un tel cas figure, qu'importe le choix de

16. Ici, comme dans le cours, quand on parle d'effet marginal de X_j sans autre précision, c'est sous-entendu l'effet marginal de X_j **sur Y**.

F, la log-vraisemblance conditionnelle serait maximisée pour $\beta_{01} \to +\infty$; intuitivement, l'échantillon dit que l'effet de X_1 est "infini". Ce n'est donc pas une histoire de multiples maxima et ce problème arrive pour un modèle probit, logit, ou tout autre choix de fonction de répartition pour F.

Question 6 (qualité du modèle, sélection des variables)

Comme le R^2 dans un modèle linéaire, le pseudo-R^2 augmente mécaniquement avec le nombre de variables explicatives (slide 30, Chapitre 3). Cela n'a rien à voir avec le fait que le modèle soit un logit, un probit, ou qu'on fasse un autre choix pour la fonction F.

On ne peut donc pas s'en servir pour sélectionner les variables à inclure dans le modèle (Alternative : voir les critères d'information de type AIC ou BIC).

Dans un modèle binaire, on peut utiliser le pseudo-R^2 pour sélectionner les variables à inclure dans le modèle

1. faux – **Vrai**, (faux est vrai ici, attention aux mots ! ;)).
2. vrai, à condition que le modèle soit un modèle logit – **Faux**.
3. vrai (quel que soit le choix de F : modèle probit, logit ou autre) – **Faux**.

Question 7 (interprétation des paramètres, sorties Stata)

Question importante, relisez les trois messages au début de la correction de la Question 4 pour l'interprétation des paramètres.

On reprend ici l'exemple du cours sur l'activité des femmes (voir la dernière Section "Application" du Chapitre 3) : $Y = 1$ si en activité (au chômage ou actif occupée), $Y = 0$ sinon.

```
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

FIGURE 1 – Sortie Stata 1

(a) On utilise ici la sortie Stata de la Figure 1. [17]

17. Rappels sur les modalités de la variable `ddipl` codant le niveau diplôme : 1 = diplôme du supérieur, c'est-à-dire, bac + 3 ou plus ; il n'y a pas de modalité 2 pour cette variable ; 3 = bac + 2 ; 4 = bac ou brevet

1. Que représente cette sortie ?

   C'est la sortie Stata de la commande `probit` : les résultats de l'estimation d'un probit. Au début, on a la valeur de la log-vraisemblance conditionnelle pour les différentes itérations de Newton-Raphson, puis sa valeur au maximum.

   A droite, au-dessus du tableau, on lit le nombre d'observations, le pseudo-R^2 et la statistique de test et la p-valeur du test de la nullité jointe de tous les coefficients (hormis la constante). Ici, cette dernière est quasiment nulle : les variables explicatives sont bien jointement utiles pour expliquer ou du moins prédire plutôt l'activité des femmes.

   Le tableau montre, pour chacun des régresseurs, le coefficient $\widehat{\beta}_j$, l'estimateur probit de β_{0j} (partie *estimation*). Pour l'*inférence*, on utilise la normalité asymptotique de l'estimateur :

   — standard errors = erreurs-types = estimation des écarts-types (sous-entendu, asymptotiques) $\longrightarrow$ pour le j-ème régresseur, c'est le j-ème terme diagonal de $\widehat{\mathcal{I}_1(\beta_0)}$ (slide 24) divisé par $\sqrt{n}$.

   — pour les tests, on a la possibilité de faire les trois tests classiques liés à l'EMV : Wald, score, ou rapport de vraisemblance (slide 27). Stata fait celui de Wald je crois (à vérifier dans l'aide mais qu'importe asymptotiquement) et montre la statistique de test (colonne `z`) et la p-valeur (colonne `P>|z|`) du test simple bilatéral de nullité du coefficient : $H_0 : \beta_{0j} = 0$ contre $H_1 : \beta_{0j} \neq 0$. On peut donc lire la significativité statistique des variables exactement de la même manière que dans une sortie de régression linéaire Stata. Par ailleurs, comme dans les régressions linéaires, on peut déterminer la significativité statistique aux seuils habituels à partir du coefficient et de l'erreur-type seulement (si pas d'accès à la p-valeur) de la façon suivante : t-stat = coefficient divisé par l'erreur-type :

      — > 1.645 : coefficient significatif à 10%
      (1.645 = quantile d'une $\mathcal{N}(0,1)$ à l'ordre $0.95 = 1 - \alpha/2$ pour $\alpha = 0.10$)

      — $> 1.960 \approx 2$: coefficient significatif à 5%
      (1.960 = quantile d'une $\mathcal{N}(0,1)$ à l'ordre $0.975 = 1 - \alpha/2$ pour $\alpha = 0.05$)

      — > 2.576 : coefficient significatif à 1%
      (2.576 = quantile d'une $\mathcal{N}(0,1)$ à l'ordre $0.995 = 1 - \alpha/2$ pour $\alpha = 0.01$)

   — Enfin, les intervalles de confiance de niveau asymptotique 95% sont obtenus à partir de la normalité asymptotique de l'estimateur probit, de l'estimateur $\widehat{\mathcal{I}_1(\beta_0)}$ consistant de la variance asymptotique et par application du lemme de Slutsky.

2. Pourquoi la modalité 1 de la variable `ddipl` n'est pas présente dans les résultats ? Reliez votre réponse à une des conditions du théorème d'identification du Chapitre 3 (slide 17).

   Pour inclure une variable explicative catégorielle, on inclut les variables indicatrices de toutes ses modalités *sauf une* (Chapitre 1, slide 3, 3ème point). Si le modèle a une constante, il faut en effet exclure une des modalités sinon il y a colinéarité parfaite ("dummy variable trap" : la somme des indicatrices de chacune des modalités vaut 1 $\longrightarrow$ colinéarité parfaite avec la constante).

   C'est la condition $\mathbb{E}[XX']$ inversible qui se rapporte à cette hypothèse d'absence de colinéarité parfaite entre les régresseurs.

3. Quelles sont les variables significatives à 1% dans ce modèle ?

   On lit cette information directement avec les p-valeurs (colonne `P>|z|`). Toutes les variables sont ici significatives à 1% exceptée la modalité numéro 3 de `ddipl` dont la p-valeur

professionnel ou autre diplôme de niveau équivalent ; 5 = CAP, BEP ou autre diplôme de niveau équivalent ; 6 = brevet des collèges, 7 = aucun diplôme (ou certificat d'étude primaire uniquement).

vaut $0.057 = 5.7\% > 5\% > 1\%$: cette variable n'est pas significative à 1%, ni à 5%, mais elle l'est à 10%.

4. *Les questions suivantes sont des vrai ou faux.*
   A l'aide de cette sortie, on peut dire que, toutes choses égales par ailleurs, [18]

   (a) être en couple par rapport à ne pas être en couple diminue la probabilité d'être en activité de 0.125. – **Faux**, le coefficient estimé pour la variable "être en couple" vaut 0.125. Mais, **attention**, on ne peut interpréter *quantitativement* ce coefficient en termes d'effet sur $\mathbb{P}(Y = 1 \,|\, X)$!

   (b) être en couple par rapport à ne pas être en couple diminue la probabilité d'être en activité de 12.5 points de pourcentage. [19] – **Faux**, idem. Remarque : 12.5 points de pourcentages (p.p.) = 0.125. [20]

   (c) être en couple par rapport à ne pas être en couple diminue la probabilité d'être en activité de 12.5%. – **Faux**, doublement faux même au sens où (i) on ne peut faire d'interprétation quantitative et (ii) on évoque un changement relatif ici et non absolu.

   (d) être en couple par rapport à ne pas être en couple diminue la probabilité d'être en activité, mais on ne peut pas avoir d'interprétation *quantitative* avec cette seule sortie, seulement *qualitative*. – **Vrai**, l'estimation de β_{0j} est négative et la variable est significative : être en couple par rapport à ne pas être en couple diminue ($\beta_{0j} < 0$) la probabilité d'être en activité ($Y = 1$).

   (e) avoir un diplôme inférieur strictement à bac + 3 (`ddipl` $\neq 1$) diminue (interprétation *qualitative* seulement) la probabilité d'être en activité par rapport à avoir un diplôme du supérieur (`ddipl` $= 1$). – **Vrai**. Attention pour les variables explicatives catégorielles, l'interprétation se fait par rapport à la modalité de référence, qui est ici `ddipl = 1` : diplôme du supérieur. Avoir un diplôme inférieur à bac + 3 contient toutes les autres modalités de `ddipl`. Les coefficients estimés de chacune de ces modalités sont négatifs et significatifs (à 10% du moins pour `ddipl = 3`) : ok pour l'interprétation qualitative : signe négatif $\longrightarrow$ diminue la probabilité $\mathbb{P}(Y = 1 \,|\, X)$.

   (f) l'effet sur la probabilité d'être en activité d'avoir un enfant supplémentaire de moins de trois ans est *qualitativement* le même que l'effet de n'avoir aucun diplôme (`ddipl` $= 7$) par rapport à avoir un diplôme du supérieur (`ddipl` $= 1$). – **Vrai**, les estimations pour les coefficients des variables `NBENF3` et `ddipl = 7` sont les deux négatives et significatives : même effet qualitatif : signe négatif $\longrightarrow$ diminue la probabilité d'être en activité.

   (g) l'effet sur la probabilité d'être en activité d'avoir un enfant supplémentaire de moins de trois ans est *quantitativement* le même (approximativement) que l'effet de n'avoir aucun diplôme (`ddipl` $= 7$) par rapport à avoir un diplôme du supérieur (`ddipl` $= 1$). – **Vrai**, oui car on peut interpréter quantitativement en termes d'effets marginaux le *ratio* des coefficients. Et ici les deux coefficients estimés sont très proches : -0.81 pour `NBENF3` et -0.80 pour `ddipl = 7`. On ne peut pas interpréter quantitativement ce -0.8 mais par contre on sait que ces deux variables ont le même effet quantitatif.

18. On admet pour formuler simplement ces questions qu'on peut avoir les interprétations causales qui suivent. En l'occurrence c'est peu crédible ici (pourquoi ?). On pourrait toujours de façon valide formuler les questions en termes de prédiction (mais les formulations sont déjà suffisamment compliquées !).

19. Un point de pourcentage (p.p.) est une unité : 1 p.p = 0.01 ; elle est généralement utilisée pour comparer des pourcentages – `https://fr.wikipedia.org/wiki/Point_de_pourcentage`.

20. On ne dit ni plus ni moins en disant, par exemple, que 10 centilitres = 1 décilitre ; c'est juste une autre unité mais utilisée pour parler d'une variation absolue et non relative entre deux pourcentages

```
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.

FIGURE 2 – Sortie Stata 2

(b) On regarde maintenant la sortie Stata de la Figure 2, obtenue après l'estimation précédente de la Figure 1.

1. Que représente cette sortie ?

   C'est la sortie Stata des effets marginaux moyens (voir première ligne) obtenue par la commande `margins, dydx(_all)` post-estimation de la commande `probit`. Les options signifient qu'on veut l'effet de X sur Y (`dydx`) pour chacune des variables explicatives (`_all`) du modèle.
   La première colonne (`dy/dx`) reporte l'estimation de l'effet marginal moyen pour chacune des variables. On a ensuite pour l'inférence les erreurs-types, la statistique de test et la p-valeur du test simple bilatéral de nullité de cet effet marginal moyen contre l'alternative, puis un intervalle de confiance de niveau asymptotique 95%. Pour l'inférence, on se sert de la normalité asymptotique de l'estimateur probit et de la delta-méthode (on estime en effet l'effet marginal moyen via une fonction de l'estimateur $\widehat{\beta}_j$, asymptotiquement normal, et de moyennes empiriques, également asymptotiquement normales).
2. A l'aide seulement du tableau de la Figure 1, était-il possible de savoir en avance le signe du coefficient estimé, ici négatif (-0.013), pour la variable `age` ?
3. Même question pour la variable `NBENF3` : à l'aide seulement du tableau de la Figure 1, était-il possible de savoir en avance le signe du coefficient estimé, ici négatif (-0.20) ?

   (*Il s'agit de deux questions plus avancées, moins importantes en première lecture par rapport aux autres, à l'exception toutefois de la façon d'estimer l'effet marginal moyen : voir les formules et explications mises en valeur en rouge ci-dessous*).

   `NBENF3` intervient simplement dans le modèle, sans puissance ni en interaction avec d'autres variables. Si β_{0j} est le coefficient devant $X_j =$ `NBENF3`, l'effet marginal de `NBENF3` vaut, en un x donné (les valeurs des différentes variables explicatives), $\phi(x'\beta_0)\beta_{0j}$, avec ϕ la densité d'une loi normale centrée réduite et l'effet marginal moyen est égal à $\mathbb{E}[\phi(X'\beta_0)]\beta_{0j}$. ϕ étant à valeurs positives, le signe de l'effet marginal moyen est égal au signe de β_{0j}. D'après la Sortie 1, l'estimée de β_{0j} est $\widehat{\beta}_j = -0.2$, négative et assez précise (voir l'intervalle de confiance). On peut donc s'attendre à un signe également négatif pour l'estimation de l'effet marginal moyen puisqu'il va être estimé en : (i) remplaçant les paramètres inconnus β_0 par leurs estimateurs probit, (ii) remplaçant l'espérance théorique inconnue par

la moyenne empirique correspondante :

$$\text{estimateur de l'effet marginal moyen } \mathbb{E}[\phi(X'\beta_0)]\beta_{0j} : \left(\frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i'\widehat{\beta}\big)\right) \times \widehat{\beta}_j,$$

où $\widehat{\beta}$ est l'estimateur probit de β_0 et $\widehat{\beta}_j$ sa j-ème composante. La fonction ϕ étant positive, cet estimateur a le même signe que l'estimateur $\widehat{\beta}_j$ et on pouvait donc bien s'attendre à ce signe.

La variable explicative `age` intervient par contre au carré dans le modèle (interaction avec elle-même). A la Sortie 1, on a estimé un coefficient positif pour le terme d'ordre 1 (0.21) et un coefficient faiblement négatif (-0.003) pour le terme d'ordre 2 : l'effet de l'âge est ainsi concave : positif puis négatif au bout d'un certain âge (point d'extremum d'un trinôme ax^2+bx+c en $-b/2a$, soit ici $\approx -0.208/(2\times -0.00284) \approx 36,6$ ans). Si β_{0l} est le coefficient devant `age`, et β_{0k} le coefficient devant la variable `age` au carré, l'effet marginal de l'âge en un $x = (x_{-\text{âge}}, x_{\text{âge}})$ donné est égal à

$$\phi(x'\beta_0)(\beta_{0l} + 2x_{\text{âge}}\beta_{0k})$$

L'effet marginal moyen est alors

$$\mathbb{E}[\phi(X'\beta_0)(\beta_{0l} + 2X_{\text{âge}}\beta_{0k})] = \mathbb{E}[\phi(X'\beta_0)]\beta_{0l} + 2\beta_{0k}\mathbb{E}[\phi(X'\beta_0)X_{\text{âge}}].$$

Pour estimer cet effet marginal moyen, on fait comme précédemment : on remplace (i) les paramètres inconnus par leurs estimateurs et (ii) les espérances théoriques par les moyennes empiriques correspondantes. On va donc utiliser l'estimateur :

$$\left(\frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i\widehat{\beta}\big)\right)\widehat{\beta}_l + 2\widehat{\beta}_k\left(\frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i\widehat{\beta}\big)X_{i,\text{âge}}\right),$$

où $X_{i,\text{âge}}$ est l'âge de la i-ème observation. Le premier terme de cette somme a le même signe que $\widehat{\beta}_l$ (puisque ϕ est positive) mais on ne peut pas connaître le signe du deuxième terme avec seulement la Sortie 1 : cela dépend de la moyenne empirique dans la population et donc de la composition de la population d'intérêt étudiée.

4. Écrivez une phrase en français pour interpréter le coefficient -0.12 (variable `ddipl` $= 6$) de cette sortie. Est-il possible d'avoir une interprétation quantitative de cette estimée ?

   Ici en Sortie 2, avec les effets marginaux moyens, on peut bien avoir une interprétation **quantitative**. Complication ici : la variable explicative considérée est catégorielle : il faut faire attention à interpréter le coefficient par rapport à la modalité de référence (qui est ici `ddipl` $= 1$, diplôme du supérieur).

   On a un effet marginal moyen estimé égal à -0.12, négatif. On l'interprète donc ainsi : toutes choses égales par ailleurs, en moyenne dans la population étudiée, l'effet d'avoir comme plus haut diplôme le brevet des collèges (`ddipl` $= 6$) par rapport à avoir un diplôme du supérieur diminue la probabilité d'être en activité de 0.12. De façon équivalente, on peut faire la même phrase en finissant par : diminue la probabilité d'être en activité de 12 points de pourcentage (c'est juste une autre unité).

5. *Les questions suivantes sont des vrai ou faux.*
   A l'aide de cette sortie, on peut dire que,[18] toutes choses égales par ailleurs, en moyenne sur la population des femmes considérée ici, avoir un enfant supplémentaire de moins de trois ans

On a estimé pour l'effet marginal moyen d'avoir un enfant de moins de trois ans en plus sur la probabilité d'être en activité : -0.20, négatif, significatif à tout niveau usuel. Ainsi, toutes choses égales par ailleurs, avoir un enfant de moins de trois ans supplémentaire diminue, en moyenne parmi la population étudiée, la probabilité d'être en activité de 0.20 = 20 points de pourcentages.

On donne parfois également une interprétation "universelle" de cet effet marginal moyen : si l'effet était "appliqué à tous les individus" (slide 13, premier point). Cela donnerait ici : si toutes les femmes de la population étudiée avaient un enfant supplémentaire de moins de trois ans (et que leurs autres caractéristiques restaient identiques – toutes choses égales par ailleurs), alors le taux d'activité des femmes dans cette population (la moyenne sur la population de la probabilité d'être en activité) diminuerait de 0.20 = diminuerait de 20 points de pourcentage.

(a) diminue la probabilité d'être en activité, et on ne peut avoir que cette interprétation *qualitative*, il n'y a pas d'interprétation *quantitative* possible ici. – **Faux**, on peut avoir une interprétation *quantitative* ici car ce sont bien des effets marginaux (ici moyens) et non les coefficients du modèle binaire.

(b) diminue la probabilité d'être en activité de 0.20. – **Vrai**.

(c) diminue la probabilité d'être en activité de 20 points de pourcentage. – **Vrai**, c'est dire la même chose que la proposition précédente.

(d) diminue la probabilité d'être en activité de 20%. – **Faux**. Attention à la différence entre une variation absolue (points de pourcentage) et une variation relative (pourcent). Exemple :

Si la probabilité de départ est égale à 0.5, une augmentation de 0.20 = une augmentation de 20 p.p. conduit à une probabilité d'arrivée égale à $0.5 + 0.2 = 0.7$ (variation, changement *absolu*). Par contre, une augmentation de 20% (variation, changement *relatif*) conduit à la probabilité d'arrivée = $0.5 \times (1 + 0.20) = 0.6$.

Question 8 (modèles statistiques)

Cette question est un peu au-delà (ou du moins à côté du cours d'Économétrie 2), c'est davantage une question de statistiques que d'économétrie en un sens. Cela peut néanmoins être intéressant pour vous mais, attention, ce n'est pas la priorité par rapport aux questions précédentes. Par ailleurs, les formulations sont peut-être rapides par moment, n'hésitez pas à me contacter[21] *en cas d'incompréhensions si vous êtes intéressés par cette question.*

(a) Retour aux modèles linéaires Relisez le cours de Statistique 1 sur ce qu'est un modèle statistique. On considère le modèle linéaire du Chapitre 1 :

$$Y = X'\beta_0 + \varepsilon$$

avec $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[X\varepsilon] = 0$. On observe un échantillon i.i.d. $(Y_i, X_i)_{i=1,\dots,n} \sim (Y, X)$.

1. Essayer d'écrire le modèle statistique associé à ce problème : le modèle *inconditionnel* (pour la loi jointe de (X_i, Y_i)). Ce modèle est-il paramétrique ? semi-paramétrique ?
2. Mêmes questions pour le modèle *conditionnel* aux régresseurs X.

Un modèle statistique est une famille de distribution des observations indexée par quelque chose, qu'on appelle le *paramètre* du modèle. Lorsque les observations sont supposées i.i.d., comme ici, il

21. Lucas Girard, adresse email Ensae habituelle : prénom.nom@ensae.fr ou via Teams.

suffit de s'intéresser à la distribution d'une observation, ici donc à la loi jointe de (Y, X), qu'on notera $\mathrm{P}^{(Y,X)}$. Un modèle statistique est alors :

$$\{P_\theta, \theta \in \Theta\}$$

où pour tout $\theta \in \Theta$, P_θ est une distribution pour le couple (Y, X), donc ici une distribution à valeurs dans $\mathbb{R} \times \mathbb{R}^K$ (si X est un vecteur aléatoire de dimension K).

Poser un modèle statistique pour ces observations, c'est supposer qu'il existe un $\theta_0 \in \Theta$ tel que P_{θ_0} est effectivement la distribution des variables aléatoires (Y, X) dont on observe des réalisations i.i.d.

θ peut être un objet (très) compliqué, infini-dimensionnel et on ne s'intéresse pas forcément à l'intégralité de θ mais à un certain paramètre $\psi(\theta)$, le *paramètre d'intérêt*. On distingue alors des modèles statistiques :

- *paramétriques* si $\Theta \subset \mathbb{R}^d$ et $\psi(\theta)$ est également fini-dimensionnel : $\psi(\theta) \in \mathbb{R}^p$
- *semi-paramétriques* si $\Theta \not\subset \mathbb{R}^d$ mais $\psi(\theta) \in \mathbb{R}^p$
- *non-paramétriques* si $\Theta \not\subset \mathbb{R}^d$ et $\psi(\theta) \notin \mathbb{R}^p$

Pour le modèle linéaire suivant

$$Y = X'\beta_0 + \varepsilon, \text{ avec } \mathbb{E}[\|X\|^2] < +\infty, \mathbb{E}[\varepsilon^2] < +\infty, \mathbb{E}[\varepsilon] = \mathbb{E}[X\varepsilon] = 0$$

on a

$$\theta = \left(\beta, \mathrm{P}^{(X,\varepsilon)}\right), \quad \Theta = \mathbb{R}^K \times \mathcal{H}, \quad \psi(\theta) = \beta,$$

où $\mathcal{H}$ est l'ensemble des distributions jointes $\mathrm{P}^{(X,\varepsilon)}$ de (X, ε) respectant les conditions de moment supposées, c'est-à-dire telles que

- $\int (\|x\|^2 + e^2)\, \mathrm{dP}^{(X,\varepsilon)}(x, e) < +\infty$ (moment d'ordre 2 fini),
- $\int e\, \mathrm{dP}^{(X,\varepsilon)}(x, e) = 0$ (ε centré),
- $\int xe\, \mathrm{dP}^{(X,\varepsilon)}(x, e) = 0$ (X et ε non corrélés).

Le modèle linéaire inconditionnel est donc semi-paramétrique : $\mathcal{H} \not\subset \mathbb{R}^d$ mais $\beta \in \mathbb{R}^K$.

Pour le modèle conditionnel, on voudrait écrire la distribution des observations Y en fixant les régresseurs. Dans ce cas, les observations ne sont plus i.i.d. mais i.n.i.d. (indépendantes non identiquement distribuées) : elles demeurent indépendantes mais les distributions sont différentes selon la valeur de X_i. P_θ doit donc être la distribution, pour un certain paramètre θ des n observations $(Y_1, \ldots, Y_n)$. En fait ici, c'est un peu pénible à écrire dans la mesure où l'on n'a pas supposé $\varepsilon \perp\!\!\!\perp X$: pour chaque observation i, la loi de ε_i peut dépendre de X_i. Le modèle conditionnel (mais ici, sans supposer $\varepsilon \perp\!\!\!\perp X$, c'est peut-être étrange de s'intéresser au modèle conditionnel) serait donc paramétré par

$$\theta = \left(\beta, \mathrm{P}^{\varepsilon_1 \mid X_1}, \ldots, \mathrm{P}^{\varepsilon_n \mid X_n}\right), \quad \Theta = \mathbb{R}^K \times \mathcal{G}, \quad \psi(\theta) = \beta,$$

où $\mathcal{G}$ est l'ensemble des n distributions conditionnelles de ε sachant $X_1, \ldots, X_n$ (il suffirait de spécifier la loi de de $\varepsilon \mid X = x$ pour tout x) vérifiant les conditions de moment. **Le point à retenir surtout est que, par rapport au cas suivant (b), le modèle conditionnel demeure ici semi-paramétrique.**

(b) EMV et MCO Dans le modèle linéaire précédent, on fait désormais l'hypothèse suivante : $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2)$.

1. Écrivez le modèle statistique *conditionnel* aux régresseurs X. Le modèle est-il paramétrique ?
2. Écrivez la vraisemblance *conditionnelle* pour une observation Y_i (à X_i fixé).
3. En déduire l'EMV de β_0 sous cette hypothèse ?
4. Quel lien a-t-il avec l'estimateur MCO de Y sur X ?

On suppose désormais $\varepsilon \perp\!\!\!\perp X$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Le modèle inconditionnel est alors seulement paramétré par :

$$\theta = (\beta, \mathrm{P}^X, \sigma^2), \quad \Theta = \mathbb{R}^K \times \mathcal{T} \times \mathbb{R}_+^*, \quad \psi(\theta) = \beta,$$

où $\mathcal{T}$ est l'ensemble des distributions P^X de X respectant la condition de moment (moment d'ordre 2 fini) pour les régresseurs : $\int \|x\|^2 \, \mathrm{dP}^X(x) < +\infty$. Ici, on a supposé ε indépendant de X et suivant une loi gaussienne centrée : la distribution de ε est donc simplement paramétrée par sa variance (qui est inconnue) σ^2.

Malgré cette simplification **le modèle inconditionnel demeure donc semi-paramétrique** car $\mathcal{T}$ est infini-dimensionnel.

Contrairement au cas (a), **le modèle conditionnel aux régresseurs est par contre paramétrique**. Sachant $X_1, \dots, X_n$, les observations $Y_1, \dots, Y_n$ sont i.n.i.d. ($Y_i = X_i'\beta_0 + \varepsilon_i$: les $(\varepsilon_i)_{i=1,\dots,n}$ sont i.i.d. mais lois différentes pour les Y_i si valeurs différentes de X_i). Le modèle conditionnel est simplement paramétré par

$$\theta = (\beta, \sigma^2), \quad \Theta = \mathbb{R}^K \times \mathbb{R}_+^*, \quad \psi(\theta) = \beta.$$

Le modèle est paramétrique est on peut donc également (alternative à l'estimateur MCO) l'estimer par maximum de vraisemblance.

Soit un paramètre $\theta = (\beta, \sigma^2)$ fixé. Conditionnellement à X_i, la loi d'une observation Y_i quelconque est une loi gaussienne (le seul aléa est ε_i variable gaussienne et la transformation affine d'une variable gaussienne est encore une variable gaussienne), d'espérance : $X_i'\beta$ et de variance σ^2 : $\mathrm{P}^{Y_i \,|\, X_i} = \mathcal{N}(X_i'\beta, \sigma^2)$. Les observations $(Y_i)_{i=1,\dots,n}$ sont indépendantes et la vraisemblance conditionnelle aux régresseurs du modèle est donc égale à

$$\mathcal{L}_n(Y_{1:n} \,|\, X_{1:n}; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2\right] = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^n \left(Y_i - X_i'\beta\right)^2\right].$$

σ^2 est ici un paramètre dit *de nuisance* qui ne nous intéresse pas. On cherche simplement à maximiser cette fonction en β. Puisque $(2\pi\sigma^2)^{-n/2}$ est une constante strictement positive, et que $z \mapsto e^{-z/(2\sigma^2)}$ est une fonction strictement décroissante, maximiser en β la vraisemblance conditionnelle revient à minimiser la fonction $\beta \mapsto \sum_{i=1}^n \left(Y_i - X_i'\beta\right)^2$. Par définition, c'est justement l'estimateur des MCO. [22]

On a donc montré que l'estimateur des MCO est égal à l'EMV dans ce modèle où l'on suppose le résidus indépendant de X et gaussien.

Voir Figures 3 à 6 ci-dessous pour quelques notes manuscrites complémentaires détaillant le raisonnement.

(c) Modèles binaires Le modèle binaire considéré dans le Chapitre 3, conditionnellement aux régresseurs, est-il paramétrique ?

La réponse est simple, c'était juste pour rendre bien explicite ce point : **oui**.

La loi des régresseurs X n'est pas spécifiée, donc le modèle inconditionnelle est semi-paramétrique. Par contre conditionnellement aux régresseurs, le modèle est bien paramétrique. Rappelez-vous que F est fixée, connue ; autrement dit, la loi du résidus ε dans le modèle linéaire sur la variable latente Y^* est totalement déterminée. Par exemple, dans un modèle probit, $\varepsilon \sim \mathcal{N}(0, 1)$ (Rappel : la variance est fixée à 1 pour pouvoir identifier le paramètre β_0)). Le modèle conditionnel est donc simplement paramétré par $\theta = \beta \in \Theta = \mathbb{R}^K$, et $\psi(\theta) = \theta = \beta$. Il n'y a même pas de paramètre de nuisance puisque la loi de ε est supposée connue.

Le modèle conditionnel étant paramétrique, on peut estimer β_0 par maximum de vraisemblance (voir la section associé du Chapitre 3). On appelle **estimateur probit** l'EMV de ce modèle lorsqu'on suppose $F = \Phi$, c'est-à-dire, $\varepsilon \sim \mathcal{N}(0, 1)$ et **estimateur logit** cet EMV lorsqu'on suppose $F = \Lambda$, c'est-à-dire lorsque ε suit une loi logistique.

22. Voir également quelques notes manuscrites complémentaires avec davantage de détails et d'étapes.

(d) Modèle de probabilité linéaire On considère le modèle de probabilité linéaire (voir la Section correspondante du Chapitre 3, slides 33 à 36) :

$$\mathbb{E}[Y \,|\, X] \overset{\text{car } Y \text{ binaire}}{=} \mathbb{P}(Y = 1 \,|\, X) \overset{\text{hypothèse}}{=} X'\beta_0.$$

1. Écrivez le modèle statistique *conditionnel* aux X.
2. Est-il paramétrique ?
3. En déduire un estimateur alternatif aux MCO de β_0 pour ce modèle.

Le but de cette question est simplement de remarquer que le modèle de probabilité linéaire est, conditionnellement aux régresseurs, également paramétrique.

On critique parfois les modèles binaires en disant qu'ils sont paramétriques (conditionnellement aux régresseurs toujours) et qu'ils obligent à spécifier une loi connue pour le résidus du modèle linéaire de la variable latente Y^* alors que les modèles linéaires sont semi-paramétriques. Mais lorsque Y est binaire, le modèle linéaire, appelé alors *modèle de probabilité linéaire*, est également paramétrique !

En effet, supposons

$$Y = X'\beta_0 + \varepsilon, \quad \mathbb{E}[\varepsilon \,|\, X] = 0$$

avec $Y \in \{0, 1\}$. On a alors

$$\mathbb{E}[Y \,|\, X] = \mathbb{P}(Y = 1 \,|\, X) = X'\beta_0.$$

Conditionnellement à $X = x$, la loi de ε est très contrainte pour assurer $Y \in \{0, 1\}$: ε est une variable aléatoire discrète

(i) prenant la valeur $1 - x'\beta_0$ avec probabilité $x'\beta_0$

(ii) et prenant la valeur $-x'\beta_0$ avec la probabilité complémentaire $1 - x'\beta_0$.

Ainsi, lorsque Y est continu et qu'on suppose

$$Y = X'\beta_0 + \varepsilon, \quad \mathbb{E}[\varepsilon \,|\, X] = 0$$

le modèle conditionnel aux régresseurs est *semi-paramétrique* : $\theta = (\beta, \mathrm{P}^{\varepsilon \,|\, X})$, la loi de ε sachant X n'est pas spécifié (on suppose juste une condition de moment conditionnel).

Par contre, lorsque Y est binaire, cette hypothèse conduit à un modèle (conditionnel aux régresseurs toujours) *paramétrique*, tout comme un modèle probit ou logit : $\theta = \beta \in \Theta = \mathbb{R}^K$. On pourrait donc également estimer β_0 dans un modèle de probabilité linéaire par maximum de vraisemblance.

FIGURE 3 – Détails Question 8 (b) – page 1

LG | ISE TP 7 2018 2019 CR3 Ex 4 partie 2 | Exercice théorique - Comparaison MCO et EMV modèle linéaire gaussien | LSE 2018 Compar. MCO EMV TP 7 Ex 4 1

$\{\varepsilon_i\}_{i=1}^n$ i.i.d

① On a $\forall i \in [\![1; n]\!]$, $n \in \mathbb{N}^*$.

$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$

i.e. $\varepsilon_i = f(X_i, Y_i)$ pour une certaine fonction f mesurable

ici $f(x,y) = y - \beta_0 - \beta_1 x$ continue car linéaire (en x et y) et donc mesurable $f: \mathbb{R}^2 \to \mathbb{R}$

les $\{Z_i = (X_i, Y_i)\}_{i=1}^n$ sont i.i.d par hypothèse,

donc les $\{\varepsilon_i\}_{i=1}^n$ sont i.i.d également en tant que fonctions mesurables des $\{Z_i\}_{i=1}^n$.

② Modèle statistique

* un modèle statistique est défini pour des observations données

↳ (étape 1) → en amont de l'écriture du modèle statistique, il faut préciser à quelles observations est associé le modèle statistique

Ici on observe d'après l'énoncé : $\{(X_i, Y_i)\}_{i=1}^n$ i.i.d, un n-échantillon bivarié.

Certes $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ mais β_0 et β_1 sont inconnus (on veut les estimer justement) donc on n'observe pas les ε_i

On écrit donc le modèle statistique pour $\{(X_i, Y_i)\}_{i=1,\dots,n}$ i.i.d.

Remarque [en économétrie ou statistique, on considère parfois un modèle conditionnel $Y_i \mid X_i$; on ne fait pas cela ici (pas la peine) ; cf détails l'année prochaine en cours de Statistique 1]

FIGURE 4 – Détails Question 8 (b) – page 2

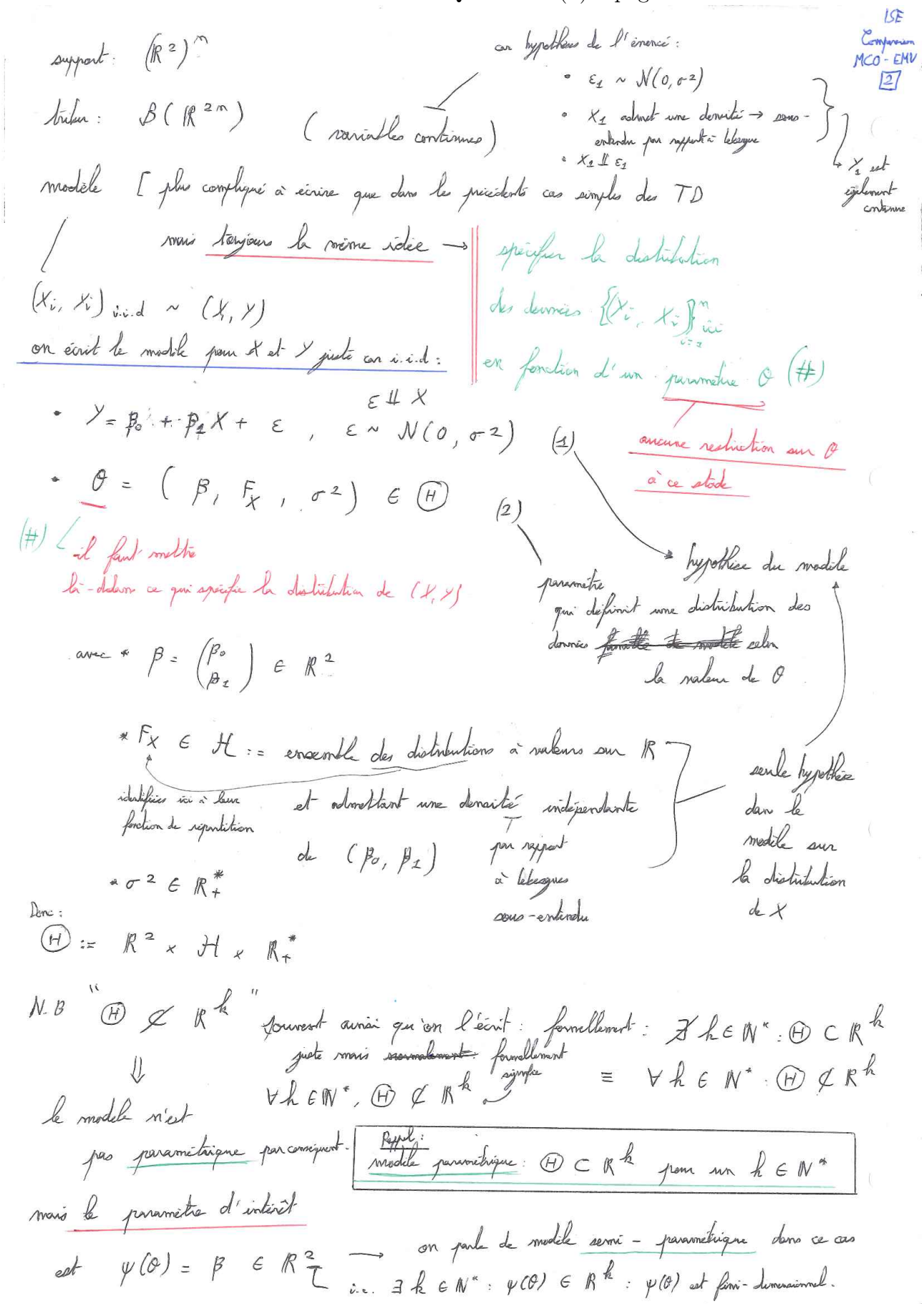

FIGURE 5 – Détails Question 8 (b) – page 3

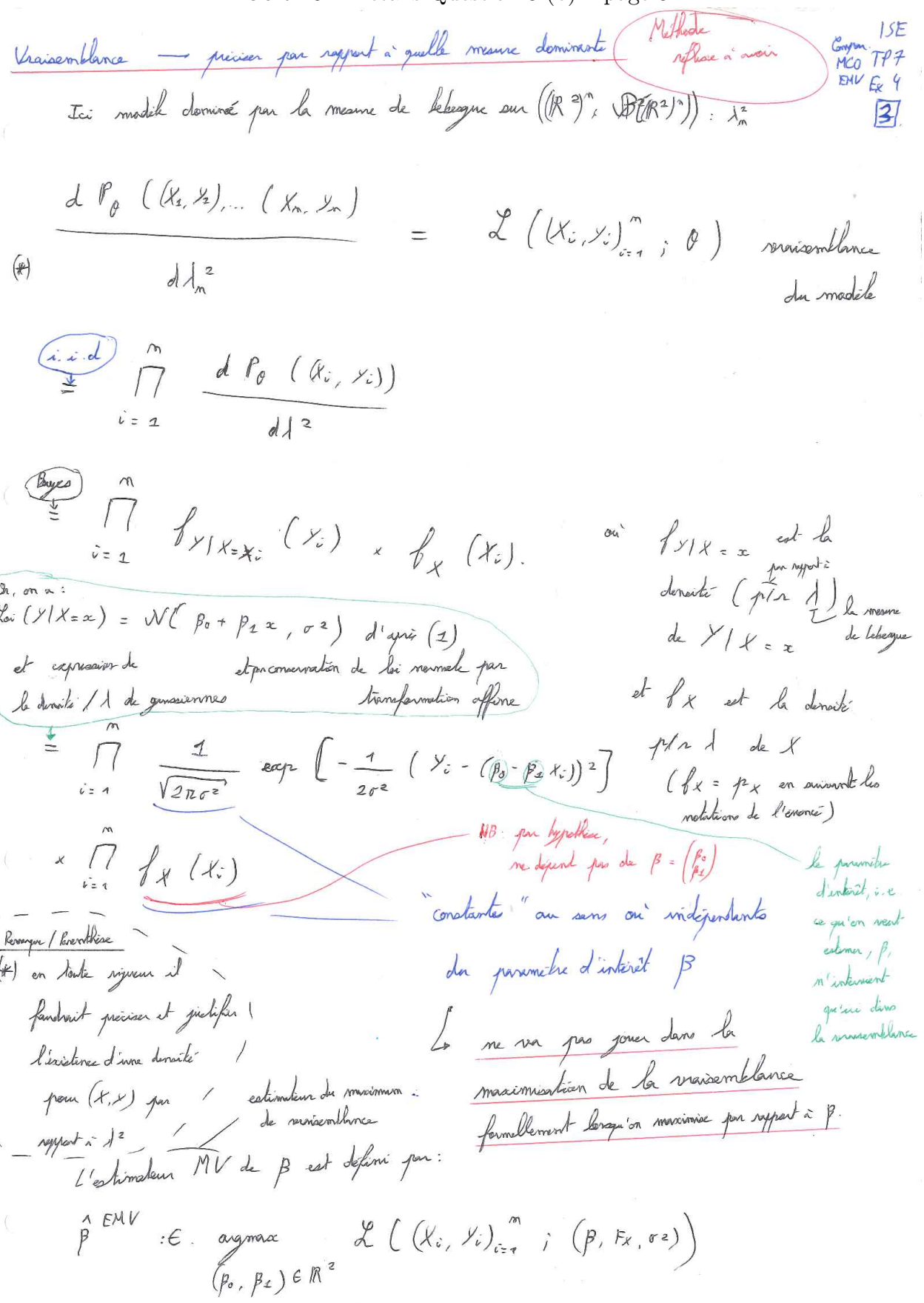

FIGURE 6 – Détails Question 8 (b) – page 4

ISE Comparaison MCO-EMV 4

a priori donc : $\hat{\beta}^{EMV} = \hat{\beta}^{EMV}\left((X_i, Y_i)_{i=1}^n, F_X, \sigma^2\right)$

fonction des observations ça c'est normal !

autre élément de θ mais $\neq$ paramètre d'intérêt → on parle de paramètre de nuisance

mais avec les hypothèses du modèle, on obtient :

seulement une fonction des observations

$$\hat{\beta}^{EMV} = \hat{\beta}^{EMV}\left((X_i, Y_i)_{i=1}^n\right) \in \underset{\beta \in \mathbb{R}^2}{\arg\max} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

soit donc :

$$\hat{\beta}^{EMV} \in \underset{\beta \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$ car $x \mapsto \exp\left(-\frac{1}{2\sigma^2} x\right)$ est décroissante.

donc comme $\hat{\beta}^{MCO}$ est défini par $\in \underset{\beta \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$

on obtient le résultat voulu : $\hat{\beta}^{EMV} = \hat{\beta}^{MCO}$.