Correction – Quiz 4 sur le Chapitre 4
(censure et sélection)

(L.G.) – Cette version : 10 avril 2022

ENSAE 2A – Économétrie 2 – Printemps 2022

Sauf mention contraire, les notations utilisées reprennent celles du cours.

Question 1 (Tobit I, identification et estimation)

On considère le modèle étudié à la première section du Chapitre 4, appelé *modèle de censure* ou *Tobit simple* ou encore *Tobit I* : on observe X et $Y = \max(0, Y^*)$ tels que

$$Y^* = X'\beta_0 + \sigma_0\varepsilon, \quad \varepsilon \,|\, X \sim \mathcal{N}(0,1).$$

Alors

1. Il est nécessaire de normaliser σ_0, typiquement à 1, pour identifier les coefficients β_0

   – **Faux**, dans un modèle Tobit I, on peut identifier jointement le vecteur de coefficients β_0 et l'écart-type σ_0 du terme d'erreur du modèle linéaire sur Y^*. Les sorties de la commande Stata `tobit` présentent l'estimation de σ_0 à la ligne `\sigma` (pour un exemple, voir l'application à la slide 13 du Chapitre 4).

2. Le modèle autorise des résidus hétéroscédastiques, c'est-à-dire, $\mathbb{V}(\varepsilon|X)$ peut dépendre de X

   – **Faux**, comme déjà vu à plusieurs reprises dans les modèles non-linéaires (modèles binaires du Chapitre 3), l'hypothèse d'indépendance entre le résidus ε et les régresseurs X implique a fortiori que les résidus sont homoscédastiques : $\mathbb{V}(\varepsilon|X) =$ constante ne dépendant pas de X :

   $$\varepsilon \perp\!\!\!\perp X \implies \mathbb{V}(\varepsilon|X) = \mathbb{V}(\varepsilon), \text{ c'est-à-dire } \varepsilon \text{ homoscédastique (par rapport à } X).$$

3. Conditionnellement aux régresseurs X, le modèle statistique est semi-paramétrique ; on ne peut pas l'estimer par maximum de vraisemblance sans ajouter d'hypothèses

   – **Faux**, on observe un échantillon i.i.d. (hypothèse implicite classique d'échantillonnage i.i.d.) des variables X et Y. Conditionnellement aux régresseurs, un modèle statistique doit donc spécifier la loi de Y sachant X. Or, $Y = \max(0, X'\beta_0 + \sigma_0\varepsilon)$ (conditionnellement à X, la seule source d'aléa dans Y vient de ε) et la loi de ε conditionnellement à X (mais on suppose ε indépendant de X de toute manière) est entièrement spécifiée (c'est une $\mathcal{N}(0,1)$).
   Le modèle statistique conditionnel aux régresseurs est donc uniquement paramétré par

   $$\theta = (\beta, \sigma) \in \Theta = (\mathbb{R}^d, \mathbb{R}_+^*),$$

   où d est la dimension de X. Le paramètre est bien fini-dimensionnel : **le modèle conditionnel aux régresseurs est paramétrique**.

4. Conditionnellement aux régresseurs X, le modèle statistique est paramétrique ; on peut l'estimer par maximum de vraisemblance

   – **Vrai**, c'est justement cet estimateur du maximum de vraisemblance qu'on appelle « estimateur Tobit (I) [1] de Y sur X » (voir slides 9 et 10 du Chapitre 4 sur l'estimation du modèle).

Question 2 (interprétation des coefficients)

Le Chapitre 4 présente deux modèles :

— le modèle *Tobit I*, aussi appelé *Tobit simple* ou *modèle de censure* (première section) ;

— le modèle *Tobit II*, aussi appelé *Tobit généralisé* ou *modèle de sélection généralisée*, parfois *modèle de sélection* simplement (deuxième section).

Cette question porte sur les paramètres d'intérêts dans ces deux modèles, plus précisément sur l'interprétation du vecteur des coefficients β_0.

Dans les deux modèles, β_0 est le coefficient des variables explicatives X dans le modèle linéaire sur Y^* (mêmes notations que dans le cours) :

$$Y^* = X'\beta_0 + \varepsilon.$$

Cette réponse anticipe également la réponse à la question 3.

Tobit II Dans le cas d'un Tobit II ou modèle de sélection (généralisée), on s'intéresse à une variable continue quantitative Y^* mais on ne l'observe pas toujours : on observe bien Y^* si $D = 1$ mais on ne l'observe pas si $D = 0$. On pose alors la variable observée Y à 0 : $Y = DY^* = Y^*\mathbb{1}\{D = 1\}$; il faut remarquer que ce choix $Y = 0$ lorsque $D = 0$ est arbitraire, cela signifie uniquement qu'on n'observe *pas* Y^* dans ce cas.

Tobit II : il y a un problème d'observations des données, on n'observe pas Y^* mais la variable d'intérêt est bien Y^*.

Par conséquent, les paramètres d'intérêt sont les effets marginaux des variables explicatives sur Y^* et le coefficient β_0, qui est bien le coefficient du modèle linéaire sur Y^* $\longrightarrow$ **On peut bien toujours interpréter quantitativement β_0 dans un modèle Tobit II**.

Tobit I Par contre, il faut **distinguer deux situations dans les modèles Tobit I ou modèle de censure** (voir notamment slides 4 et 5 du Chapitre 4) :

(a) **Cas de *données censurées* en raison d'un problème d'observation des données : la véritable variable d'intérêt est alors Y^*** continue quantitative, mais on ne l'observe que dans certains cas (au-dessus ou en-dessous d'un certain seuil) $\longrightarrow$ on peut interpréter quantitativement β_0 sur la variable d'intérêt qui est Y^*.

(b) **Cas de *solutions en coin* : la variable d'intérêt est Y** alors que Y^* est une variable latente potentiellement dépourvue de sens quantitatif précis. Typiquement, Y^* représente l'utilité de consommer un bien (utilité exprimée en nombre d'unités qu'on souhaiterait idéalement consommer, et donc potentiellement un nombre négatif) alors que Y est le nombre d'unités effectivement consommées de ce bien. Y a bien un sens quantitatif dans ce cas mais les coefficients β_0 concernent Y^* qui n'a pas de sens quantitatif précis $\longrightarrow$ on ne peut pas interpréter quantitativement β_0 sur la variable d'intérêt qui est Y, il faut passer par les effets marginaux (voir question suivante).

1. Parfois sans préciser Tobit I car si on parle seulement de Y et de X sans évoquer une variable de sélection D, on ne peut pas être dans un modèle Tobit II.

Ainsi, dans le modèle Tobit I, on peut interpréter directement quantitativement β_0 *seulement dans le cas de données censurées*.

L'interprétation *quantitative* directe du vecteur de coefficients β_0

1. n'est jamais possible dans le cas d'un modèle Tobit I – **Faux**, c'est possible dans le cas "données censurées".
2. est toujours possible dans le cas d'un modèle Tobit I – **Faux**, cela n'est *pas* possible dans le cas "solutions en coin".
3. n'est jamais possible dans le cas d'un modèle Tobit II – **Faux**.
4. est toujours possible dans le cas d'un modèle Tobit II – **Vrai**.

Question 3 (variable expliquée d'intérêt et interprétation)

On peut distinguer trois situations dans le Chapitre 4 :

(i) le modèle Tobit I dans le cas de données censurées

(ii) le modèle Tobit I dans le cas de solutions en coin

(iii) le modèle Tobit II

Pour chacune de ces trois situations,

1. Précisez quelle est la variable d'intérêt : Y ou Y^* ?
2. Pour une variable explicative continue X_k, dans le cas "simple" où cette variable explicative intervient sans interaction ou puissance dans le modèle, indiquez le ou les paramètres d'intérêt classiques, c'est-à-dire les effets marginaux de X_k sur ... ?

(i) Modèle Tobit I dans le cas de données censurées

1. La variable d'intérêt est Y^* (voir réponse à la question précédente).
2. Le paramètre d'intérêt est [2] l'effet marginal de X_k sur la variable d'intérêt qui est Y^* :

$$\frac{\partial \mathbb{E}[Y^* \,|\, X_k = x_k, X_{-k} = x_{-k}]}{\partial x_k}$$

qui vaut simplement β_{0k} lorsque la variable explicative X_k apparaît "simplement" dans le modèle. C'est justement pour cela qu'on *peut bien* interpréter directement quantitativement les coefficients β_0 dans cette situation de Tobit I avec "données censurées".

(ii) Modèle Tobit I dans le cas de solutions en coin

1. La variable d'intérêt est Y (voir réponse à la question précédente).
2. Pour une variable explicative X_k donnée, on peut considérer plusieurs paramètres d'intérêt en distinguant la marge *extensive* ($Y = 0$ ou $Y > 0$?) et la marge *intensive* (sachant $Y > 0$, quelle valeur de Y ?). Plus précisément, on s'intéresse à

$$\frac{\partial \mathbb{E}[Y \,|\, X_k = x_k, X_{-k} = x_{-k}]}{\partial x_k} = \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)\beta_{0k},$$

2. Ou plutôt, les paramètres d'intérêt sont, car il s'agit en général d'une *fonction* d'effet marginal dès lors que X_k intervient également dans le modèle avec une puissance ou en interaction avec une autre variable explicative.

l'effet marginal "total" de X_k sur Y, "total" au sens où il combine marge extensive et intensive (équation (1), slide 5 du Chapitre 4), et à

$$\frac{\partial \mathbb{E}[Y \mid Y > 0, X_k = x_k, X_{-k} = x_{-k}]}{\partial x_k} = \left[1 + \lambda'\left(\frac{x'\beta_0}{\sigma_0}\right)\right] \beta_{0k},$$

l'effet marginal de X_k sur Y sur la marge intensive uniquement (équation (2), slide 5 du Chapitre 4 et propriété : $\lambda'(x) = -\lambda(x)[x + \lambda(x)]$ pour tout réel x).

Remarque : ces deux paramètres sont *distincts* de β_{0k} et c'est précisément pourquoi on *ne peut pas* directement interpréter quantitativement les coefficients sur la variable d'intérêt qui est Y dans le cas Tobit I avec "solutions en coin".

(iii) Modèle Tobit II

1. La variable d'intérêt est Y^* (voir réponse à la question précédente).
2. Comme dans le cas Tobit I avec données censurées, le paramètre d'intérêt est l'effet marginal de X_k sur la variable d'intérêt qui est Y^* :

$$\frac{\partial \mathbb{E}[Y^* \mid X_k = x_k, X_{-k} = x_{-k}]}{\partial x_k}$$

qui est juste égal à β_{0k} lorsque la variable explicative X_k apparaît "simplement" dans le modèle, sans interaction ou puissance. A nouveau, c'est en raison de cette égalité au coefficient β_{0k} (dans le cas simple sans puissance ou interaction) qu'*on peut bien* interpréter directement quantitativement les coefficients β_0 dans un modèle Tobit II.

Question 4 (Tobit II, sélection exogène)

On considère le modèle Tobit II ou Tobit généralisé de la deuxième section du Chapitre 4 :

$$\begin{aligned} Y^* &= X'\beta_0 + \varepsilon \\ D &= \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\}. \end{aligned}$$

Dans ce modèle, donnez deux définitions équivalentes de la sélection dite *exogène*.
Remarque : la question demande des définitions formelles faisant intervenir des indépendances entre certaines variables aléatoires, peut-être conditionnellement à d'autres.

De manière générale, **la sélection exogène** signifie que, conditionnellement aux variables explicatives observées, la variable d'intérêt est indépendante de la variable de sélection (l'indicatrice d'observer la variable d'intérêt).[3] Dit autrement, conditionnellement aux variables explicatives, le fait de ne pas observer la variable d'intérêt Y^* (indicatrice d'observation $D = 0$) ne dépend pas de la valeur de Y^*. La non-observation ($D = 0$) ou, vue de façon complémentaire, la sélection ($D = 1$) est aléatoire conditionnellement aux régresseurs. En anglais, on parle parfois de "missing-at-random" pour qualifier cette situation.

Dans un modèle Tobit II, avec les notations vues en cours,

— Y^* est la variable expliquée d'intérêt ;
— D est l'indicatrice de sélection / d'observation, égale à 1 si on observe Y^*, 0 sinon ;
— X et Z sont les variables explicatives observées, qui peuvent avoir certaines composantes en commun.

3. *Il est préférable de se souvenir ainsi de l'idée générale avec des* MOTS*, et non seulement de symboles ou formules (comme par exemple, $Y^* \perp\!\!\!\perp D \mid X$) pour pouvoir les adapter à d'autres cas et d'autres notations.*

Ainsi, dans ce cadre, la sélection exogène s'écrit formellement

$$Y^* \perp\!\!\!\perp D \,|\, (X, Z).$$

Pour une deuxième écriture équivalente, on peut remarquer que

— à X fixé, seul ε est aléatoire dans Y^* car $Y^* = X'\beta_0 + \varepsilon$;

— de même, puisque $D = \mathbb{1}\{Z'\gamma_0 + \eta\}$, sachant Z, l'aléa de D provient uniquement de η.

Formellement, sachant X, Y^* est une fonction mesurable de ε et d'un terme non-stochastique β_0 (idem pour D qui est une fonction de η et du paramètre γ_0 sachant Z). Par le lemme des coalitions, on a ainsi

$$Y^* \perp\!\!\!\perp D \,|\, (X, Z) \iff \varepsilon \perp\!\!\!\perp \eta \,|\, (X, Z),$$

c'est-à-dire que, conditionnellement aux régresseurs, les deux résidus (celui impactant la variable d'intérêt et celui associé à la variable de sélection) sont indépendants.

Question 5 (Tobit I, estimation)

On considère le modèle de censure suivant

$$Y = \max(0, X'\beta_0 + \varepsilon), \quad \varepsilon \,|\, X \sim \mathcal{N}(0, \sigma_0^2).$$

On observe un échantillon i.i.d. $(Y_i, X_i)_{i=1,\dots,n} \sim (Y, X)$.
Dans ce modèle,

1. L'estimateur MCO de Y sur X (utilisant toutes les observations $i \in \{1, \dots, n\}$) est un estimateur consistant de β_0

   – **Faux**, l'estimateur des Moindres Carrés Ordinaires (MCO) de Y sur X, utilisant toutes les données, y compris les données « censurées » avec $Y = 0$, n'est pas un estimateur consistant de β_0 en général (voir premier point du slide 8 du Chapitre 4). Il approche l'effet marginal moyen de X sur Y, c'est-à-dire $\mathbb{P}(Y > 0)\,\beta_0$,[4] d'où un biais d'atténuation : l'estimateur MCO aura tendance à être plus petit en valeur absolue que le paramètre β_0 cible.

2. L'estimateur MCO de Y sur X calculé uniquement sur les observations $\{i : Y_i > 0\}$ est un estimateur consistant de β_0

   – **Faux**, l'estimateur des MCO de Y sur X utilisant seulement les observations dites « non-censurées » $\{i : Y_i > 0\}$ n'est pas un estimateur consistant de β_0 en général (voir troisième point du slide 8 du Chapitre 4). A nouveau, il aura tendance à estimer un effet marginal moyen de X sur Y, ici sachant $Y > 0$ puisqu'on se restreint aux données non-censurées. L'effet marginal moyen de X sur Y sachant $Y > 0$ vaut $\beta_0\,\{1 + \mathbb{E}[\lambda'(X'\beta_0/\sigma^2)]\}$, d'où à nouveau un biais d'atténuation vers 0 puisque pour tout réel x, $\lambda'(x) \in]-1, 0[$.

3. L'estimateur du maximum de vraisemblance de β_0 est un estimateur consistant de β_0 même si les résidus sont hétéroscédastiques (c'est-à-dire, $\mathbb{V}(\varepsilon|X)$ dépend de X)

   – **Faux**, l'EMV $\widehat{\beta}$ de β_0 sera bien un estimateur consistant et asymptotiquement normal de β_0 (résultat au dernier point de la slide 10 du Chapitre 4) *à condition que la modèle soit correctement spécifié*. Cela nécessite d'avoir $\varepsilon \perp\!\!\!\perp X$ et donc, a fortiori, des résidus ε homoscédastiques (la variance de ε conditionnellement à X ne dépend pas de X). Si les résidus sont hétéroscédastiques, le modèle est mal-spécifié et $\widehat{\beta}$ n'est pas un estimateur consistant de β_0.

4. Il s'agit d'un vecteur : les effets marginaux moyens de chacune des composantes de X sur Y.

4. L'estimateur du maximum de vraisemblance de β_0 n'utilise pas les observations censurées, c'est-à-dire les observations $\{i : Y_i = 0\}$

   – **Faux**, l'EMV utilise toutes les observations, à la fois les $\{i : Y_i > 0\}$ et les $\{i : Y_i = 0\}$. Cela est visible dans l'écriture de la vraisemblance du modèle (voir slides 9 et 10 du Chapitre 4).

5. On peut estimer de façon consistante β_0/σ_0 par un probit de $D = \mathbb{1}\{Y > 0\}$ sur X

   – **Vrai**. Posons $D := \mathbb{1}\{Y > 0\}$. On a alors

   $$D = \mathbb{1}\{X'\beta_0 + \varepsilon > 0\}, \quad \varepsilon \,|\, X \sim \mathcal{N}(0, \sigma^2).$$

   Il s'agit donc quasiment d'un modèle probit de D sur X ; quasiment car la variance du résidus n'est pas normalisée à 1 par rapport à un probit classique. C'est pourquoi un probit de D sur X ne permettra d'identifier et d'estimer de façon consistante que β_0 à une constante multiplicative près : β_0/σ_0.

   Toutefois, même si on peut faire cela, on n'a donc pas intérêt à le faire car, intuitivement, *on perd de l'information en oubliant Y pour ne considérer que D* :

   — Modèle binaire : la variable observée $D \in \{0, 1\}$ indique uniquement si la variable latente $Y^* = X'\beta_0 + \varepsilon$ est au-dessus ou en-dessous de 0 ;

   — Modèle Tobit I ou de censure : la variable observée Y donne en plus la valeur de Y^* lorsque Y^* est au-dessus de 0. C'est grâce à cette information additionnelle qu'on n'a pas besoin de normaliser la variance du résidus et qu'on peut identifier conjointement les paramètres β_0 et σ_0.

Question 6 (interprétation, sorties Stata)

On s'intéresse aux déterminants du logarithme du salaire horaire des femmes (variable `logsal_hor`). Cette variable n'est observée que pour les femmes qui sont actives et employées sur le marché du travail : variable `indic_obs` est égale à 1 si la femme travaille (on observe alors son salaire horaire), 0 sinon.

On dispose également des variables suivantes :

— le diplôme `ddipl` en six modalités : $\geq$ bac+3 (modalité 1) ; bac+2 (2) ; bac ou équivalent (3) ; CAP, BEP ou équivalent (4) ; brevet des collèges (5) ; aucun diplôme ou CEP (6) ;

— l'expérience potentielle `exp` (définie comme l'âge actuel – l'âge à la fin des études) et le carré de l'expérience potentielle `exp2` ;

— le nombre d'enfants de moins de six ans `NBENF6`.

On estime un modèle et on obtient la sortie Stata présentée en Figure 1.

Les proposition suivantes sont des vrai ou faux. Si vous pensez que la proposition est fausse, indiquez alors la bonne réponse ou expliquez pourquoi la proposition est fausse.

1. Le modèle estimé ici est un Tobit simple, aussi appelé Tobit I. – **Faux**, il s'agit ici d'un Tobit II ou modèle de sélection généralisée. C'est l'application de la deuxième section du Chapitre 4 (slides 26, 27 et 28). Avec les notations du cours, la variable expliquée d'intérêt est $Y^* =$ `logsal_hor` et la variable binaire d'observation / sélection est $D =$ `indic_obs`.

2. Les estimations présentées ont toutes été obtenues par maximum de vraisemblance. – **Faux**, les estimations ont été obtenues par la commande Stata `heckman` avec l'option `twostep` : cette option spécifie une estimation par la méthode « Heckit » en deux étapes.

```
. xi: heckman logsal_hor i.ddipl exp exp2, select(indic_obs = NBENF6 exp ///
>                         exp2 i.ddipl) twostep
i.ddipl           _Iddipl_1-6         (_Iddipl_6 for ddipl==7 omitted)

Heckman selection model -- two-step estimates   Number of obs      =       106878
(regression model with sample selection)        Censored obs       =        89296
                                                   Uncensored obs     =       17582

                                                Wald chi2(7)       =      5490.36
                                                Prob > chi2        =       0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logsal_hor						
_Iddipl_1	.6924857	.0116691	59.34	0.000	.6696146	.7153568
_Iddipl_2	.5026082	.0139363	36.06	0.000	.4752936	.5299229
_Iddipl_3	.305763	.0122888	24.88	0.000	.2816773	.3298487
_Iddipl_4	.149917	.0117749	12.73	0.000	.1268387	.1729953
_Iddipl_5	.1580453	.013888	11.38	0.000	.1308254	.1852653
exp	.0276042	.0014949	18.47	0.000	.0246742	.0305342
exp2	-.0004318	.0000382	-11.31	0.000	-.0005067	-.000357
_cons	4.276439	.0857463	49.87	0.000	4.108379	4.444499
indic_obs						
NBENF6	-.1443794	.0091964	-15.70	0.000	-.162404	-.1263548
exp	.0281243	.001419	19.82	0.000	.0253431	.0309055
exp2	-.0008314	.0000293	-28.37	0.000	-.0008888	-.000774
_Iddipl_1	.0607705	.0175224	3.47	0.001	.0264273	.0951138
_Iddipl_2	.2154777	.0171016	12.60	0.000	.1819592	.2489962
_Iddipl_3	.1455301	.0163727	8.89	0.000	.1134402	.1776201
_Iddipl_4	.1406145	.0155556	9.04	0.000	.1101261	.1711029
_Iddipl_5	.1193359	.0203689	5.86	0.000	.0794137	.1592581
_cons	-1.111414	.0205948	-53.97	0.000	-1.151779	-1.071049
mills						
lambda	.2169139	.0503266	4.31	0.000	.1182756	.3155523
rho	0.53462					
sigma	.40573343					

FIGURE 1 – Sortie Stata pour la Question 6

Sans cette option, Stata estime le modèle par maximum de vraisemblance en supposant (ε, η) gaussien (slide 24 du Chapitre 4). [5]

3. Les résultats montrent un biais de sélection négatif au sens où, toutes choses égales par ailleurs, les femmes qui ne travaillent pas auraient un salaire supérieur à celles qui travaillent. – **Faux**. Le coefficient de la ligne `rho` est un estimateur de la corrélation entre les deux résidus : ε (le résidus du modèle sur Y^*) et η (le résidus du modèle sur D). Il est ici positif (et de façon significative d'après l'inférence sur le paramètre `lambda` – voir Figures 2, 3 et 4 à la fin de ce document pour des explications plus détaillées).

   ε et η sont ainsi positivement corrélés : toutes choses égales par ailleurs (c'est-à-dire, à X et Z fixés), lorsque ε est grand (le salaire potentiel Y^* est plus élevé), η est également plus élevé : D a plus de chances d'être positif, l'individu a plus tendance à participer au marché du travail. Dit dans l'autre sens, un ε faible ou négatif (salaire faible) est associé à η faible ou négatif (plus de chances de ne pas participer au marché du travail : $D = 0$).

   Les résultats montrent donc une sélection significativement endogène mais entraînant un biais où, toutes choses égales par ailleurs, les femmes qui ne travaillent pas auraient au contraire un salaire *inférieur* à celles qui travaillent.

4. On ne peut pas interpréter directement quantitativement la valeur des coefficients dans la partie `logsal_hor` des sorties. – **Faux**, on *peut interpréter directement quantitativement*

5. Plus précisément, les estimations pour le modèle probit de D sur Z (première étape du Heckit) – coefficients de la partie `indic_obs` du tableau – sont effectivement obtenues par maximum de vraisemblance puisqu'il s'agit d'un probit ; mais ce n'est pas le cas des autres estimations.

les coefficients dans la partie `log_sal` des sorties car ce sont bien les coefficients du modèle linéaire sur Y^* qui est la variable d'intérêt (voir Questions 2 et 3 ci-dessus).

5. On ne peut pas interpréter directement quantitativement la valeur des coefficients dans la partie `indic_obs` des sorties. – **Vrai**, il s'agit ici d'un modèle probit de D sur Z et *on ne peut donc pas interpréter directement quantitativement les coefficients estimés dans la partie* `indic_obs` *des sorties* (voir Chapitre 3 et Quiz 3).

6. Toutes choses égales par ailleurs, être diplômée du supérieur (bac+3 ou davantage) augmente le salaire potentiel de 39% environ par rapport à n'avoir que le bac. – **Vrai**. Il y a trois arguments à combiner pour comprendre cela :
   (a) la réponse à la proposition 4 ci-dessus : on peut bien interpréter directement *quantitativement* les coefficients de la partie `logsal_hor` des sorties sur Y^* ;
   (b) Y^* est ici le *logarithme* du salaire horaire et les variables explicatives sont en niveau ou des indicatrices des différentes modalités d'éducation $\longrightarrow$ *modèle log-level* $\longrightarrow$ interprétation en termes de changement relatif : une variation de 1 d'un régresseur X_k entraîne (approximativement) une variation relative de $100 \times \beta_{0k}$% de la variable dont on prend le logarithme ;
   (c) enfin, on s'intéresse ici à des *variables explicatives catégorielles*. Avoir bac+3 ou davantage est la modalité 1 de `ddipl` alors qu'avoir uniquement le bac est la modalité 3. On lit sur les sorties que les coefficients estimés associés sont respectivement 0.692 et 0.306.

   Ainsi, avoir bac+3 par apport au bac uniquement entraîne une modification (approximativement) de $100 \times (0.692 - 0.306)$% du salaire horaire potentiel, soit une hausse d'environ 39%.

7. Toutes choses égales par ailleurs, avoir un enfant en plus âgé de moins de six ans diminue la probabilité d'être employée de 14% environ. – **Faux**, *on ne peut pas interpréter directement quantitativement les coefficients du modèle probit* de $D =$ `indic_obs` sur $Z = (1,$ `NBENF6`, `exp`, `exp` au carré, `ddipl`$)'$. Le coefficient estimé de `NBENF6` est -0.14 mais il n'a pas de sens quantitatif direct.

8. Toutes choses égales par ailleurs, avoir un enfant en plus âgé de moins de six ans diminue la probabilité d'être employée de 14 points de pourcentage environ. – **Faux**, même réponse (voir Quiz 3 pour la distinction entre pourcentage et point de pourcentage).

9. Avec cette sortie, on peut seulement dire que, toutes choses égales par ailleurs, avoir un enfant en plus âgé de moins de six ans diminue la probabilité d'être employée (interprétation qualitative uniquement). – **Vrai**, on ne peut interpréter que *qualitativement* les coefficients du modèle probit de $D =$ `indic_obs` sur Z. Ici, le coefficient estimé est négatif (-0.14) : toutes choses égales par ailleurs (à expérience et niveau de diplôme fixés ici), avoir plus d'enfants diminue la probabilité d'avoir $D = 1$, c'est-à-dire, d'être employée. Pour une interprétation *quantitative*, il faudrait calculer les effets marginaux sur D.

10. Le modèle estimé ici repose sur l'hypothèse que la variable `NBENF6` n'a pas d'effet direct sur `logsal_hor`. – **Vrai**, le modèle Tobit II nécessite d'avoir au moins une composante de Z exclue de X (voir slide 19 du Chapitre 4). Ici, cette composante est la variable `NBENF6` et on suppose donc que cette variable n'est pas dans le modèle sur $Y^* =$ `logsal_hor`, c'est-à-dire n'a pas d'effet direct sur le salaire horaire. En l'occurrence, on peut ici douter de la crédibilité de cette hypothèse.

Question 7 (Tobit I, interprétation et estimation)

On considère le modèle de censure suivant

$$Y = \max(0, Y^*), \quad Y^* = X'\beta_0 + \varepsilon, \quad \varepsilon \,|\, X \sim \mathcal{N}(0, \sigma_0^2).$$

On observe un échantillon i.i.d. $(Y_i, X_i)_{i=1,\dots,n} \sim (Y, X)$.
Dans ce modèle,

1. L'estimateur MCO de Y sur X (utilisant toutes les observations $i \in \{1, \dots, n\}$) est un estimateur consistant de l'effet marginal moyen de X sur Y

   – **Faux**, cela peut sembler contredire la réponse à la Question 5 mais c'est une approximation, une tendance seulement : l'estimateur MCO a *tendance* à être proche de l'effet marginal moyen de X sur Y mais ce n'est pas la limite en probabilité de l'estimateur MCO.

   Pour estimer de façon consistante l'effet marginal moyen de X_k sur Y (ce qui n'a d'intérêt que dans le cas d'un Tobit I "solution en coin", où la variable d'intérêt est bien Y, et non Y^*), il faut procéder en plusieurs étapes :

   (a) estimer de façon consistante β_{0k} par l'estimateur $\widehat{\beta}_k$ Tobit I (EMV) de Y sur X ;

   (b) estimer de façon consistante $\mathbb{P}(Y > 0)$ par la fréquence empirique $n^{-1} \sum_{i=1}^n \mathbb{1}\{Y_i > 0\}$ (méthode des moments) ;

   (c) estimer finalement l'effet marginal moyen de X_k sur Y :

   $$\mathbb{E}\left[\frac{\partial \mathbb{E}[Y \,|\, X]}{\partial X_k}\right] = \mathbb{P}(Y > 0)\, \beta_{0k}$$

   par

   $$\widehat{\beta}_k \times n^{-1} \sum_{i=1}^n \mathbb{1}\{Y_i > 0\}\,.$$

   Il s'agit bien d'un estimateur consistant de l'effet marginal moyen de X_k sur Y par le Continuous Mapping Theorem (le produit est une opération continue).

2. L'estimateur MCO de Y sur X calculé uniquement sur les observations $\{i : Y_i > 0\}$ est un estimateur consistant de l'effet marginal moyen de X sur Y

   – **Faux**, même raison que précédemment et même "doublement faux" au sens où l'estimateur MCO de Y sur X sur les données non-censurées seules approche l'effet marginal moyen de X sur Y *sachant* $Y > 0$.

3. β_0 peut s'interpréter comme l'effet marginal moyen de X sur Y^*

   – **Vrai**, β_0 est le coefficient du modèle linéaire sur Y^*. C'est bien pourquoi on peut directement interpréter *quantitativement* les coefficients β_0 lorsque la variable d'intérêt est Y^* dans un Tobit I pour le cas de "données censurées".

4. On peut estimer de façon consistante β_0/σ_0 par un logit de $D = \mathbb{1}\{Y > 0\}$ sur X – **Faux**, ce serait vrai pour un *probit* mais non un *logit* (voir Question 5). On suppose en effet le résidus gaussien dans les modèles Tobit I.

Question 8 (Tobit I, log-vraisemblance)

On considère le modèle Tobit I

$$Y = \max(0, X'\beta_0 + \varepsilon) \text{ avec } \varepsilon \,|\, X \sim \mathcal{N}(0, \sigma_0^2).$$

Dans ce modèle, la log-vraisemblance (conditionnelle aux régresseurs) correspondant à une observation (y, x) et aux paramètres (β, σ) s'écrit

La bonne réponse est la réponse 2. Il suffit de prendre le logarithme de la vraisemblance $g(y \,|\, x)$ calculée à la dernière ligne du slide 9 du Chapitre 4.

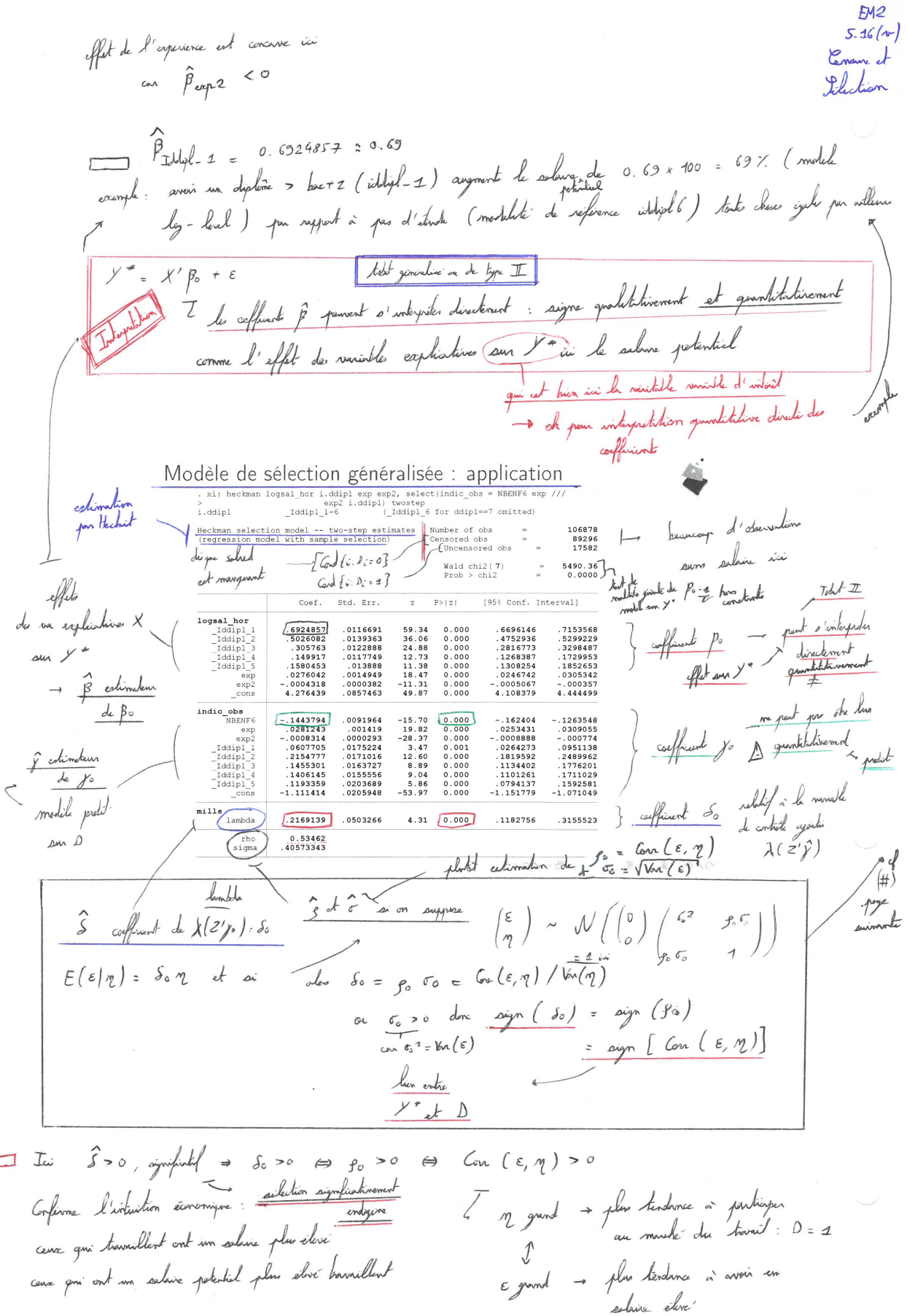
Sans reconnaître exactement cette forme, on peut aussi aboutir à ce résultat par élimination (puisqu'il n'y a qu'une seule bonne réponse) :

— la réponse 4. ne peut être vraie car ne prend pas en compte le cas $Y = 0$;

— les réponses 1. et 3. ne peuvent être vraies car la vraisemblance doit faire intervenir à la fois une densité ϕ (pour le cas continu $Y > 0$) et une probabilité / fonction de répartition Φ (pour le cas de la masse en $Y = 0$).

1. $\ln\left[\Phi\left(\frac{y-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y>0\} + \ln\left[\Phi\left(\frac{-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y=0\}$ – **Faux.**
2. $\ln\left[\frac{1}{\sigma}\,\phi\left(\frac{y-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y>0\} + \ln\left[\Phi\left(\frac{-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y=0\}$ – **Vrai.**
3. $\ln\left[\frac{1}{\sigma}\,\phi\left(\frac{y-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y>0\} + \ln\left[\frac{1}{\sigma}\,\phi\left(\frac{-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y=0\}$ – **Faux.**
4. $\ln\left[\frac{1}{\sigma}\,\phi\left(\frac{y-x'\beta}{\sigma}\right)\right]\mathbb{1}\{y>0\}$ – **Faux.**

FIGURE 2 – Commentaires et explications (sorties Stata d'un modèle Tobit II) – (1).



		Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
logsal_hor						
	_Iddipl_1	.6924857	.0116691	59.34	0.000	.6696146 .7153568
	_Iddipl_2	.5026082	.0139363	36.06	0.000	.4752936 .5299229
	_Iddipl_3	.305763	.0122888	24.88	0.000	.2816773 .3298487
	_Iddipl_4	.149917	.0117749	12.73	0.000	.1268387 .1729953
	_Iddipl_5	.1580453	.013888	11.38	0.000	.1308254 .1852653
	exp	.0276042	.0014949	18.47	0.000	.0246742 .0305342
	exp2	-.0004318	.0000382	-11.31	0.000	-.0005067 -.000357
	_cons	4.276439	.0857463	49.87	0.000	4.108379 4.444499
indic_obs						
	NBENF6	-.1443794	.0091964	-15.70	0.000	-.162404 -.1263548
	exp	.0281243	.001419	19.82	0.000	.0253431 .0309055
	exp2	-.0008314	.0000293	-28.37	0.000	-.0008888 -.000774
	_Iddipl_1	.0607705	.0175224	3.47	0.001	.0264273 .0951138
	_Iddipl_2	.2154777	.0171016	12.60	0.000	.1819592 .2489962
	_Iddipl_3	.1455301	.0163727	8.89	0.000	.1134402 .1776201
	_Iddipl_4	.1406145	.0155556	9.04	0.000	.1101261 .1711029
	_Iddipl_5	.1193359	.0203689	5.86	0.000	.0794137 .1592581
	_cons	-1.111414	.0205948	-53.97	0.000	-1.151779 -1.071049
mills						
	lambda	.2169139	.0503266	4.31	0.000	.1182756 .3155523
	rho	0.53462				
	sigma	.40573343				

FIGURE 3 – Commentaires et explications (sorties Stata d'un modèle Tobit II) – (2).

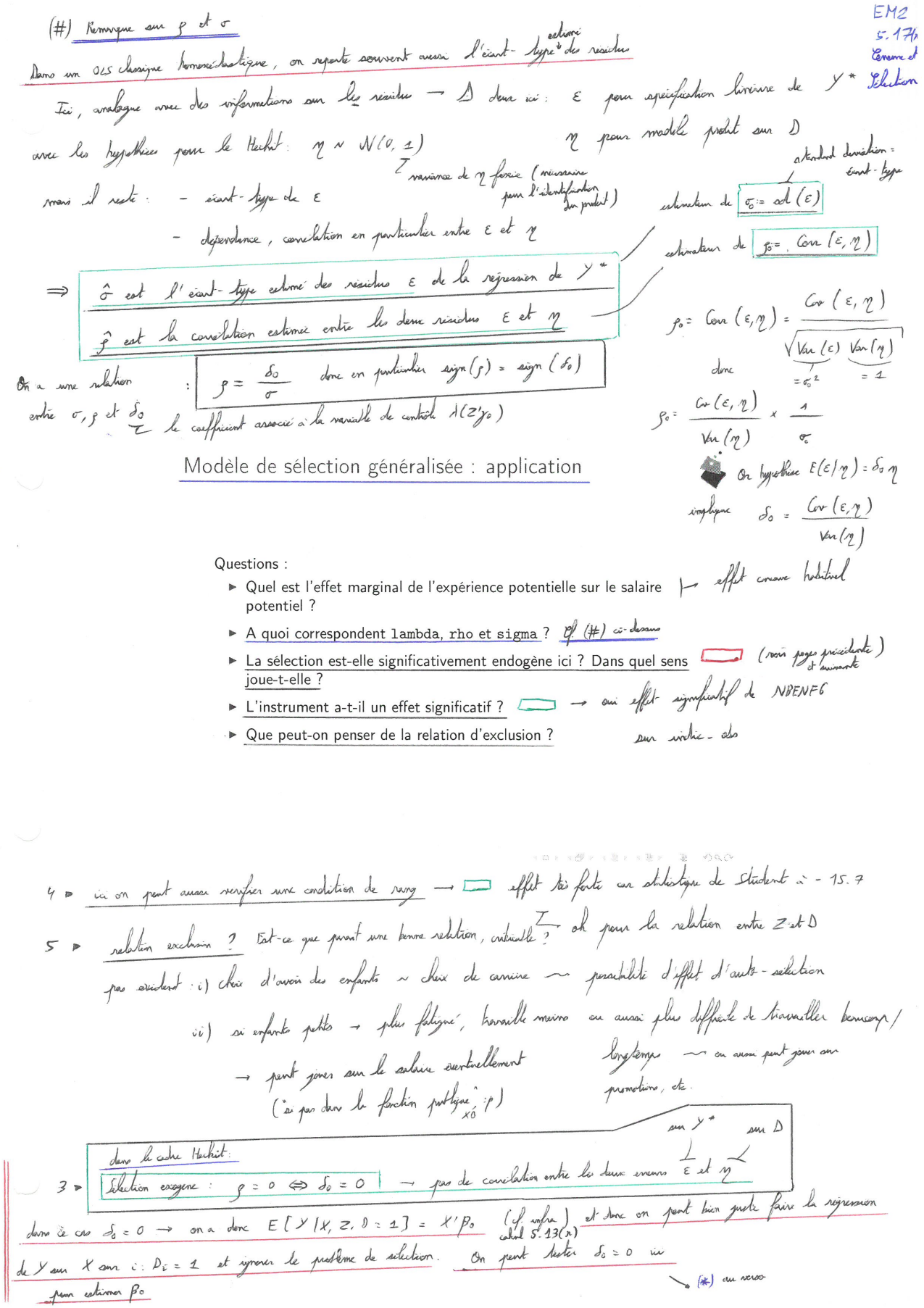


FIGURE 4 – Commentaires et explications (sorties Stata d'un modèle Tobit II) – (3).

(*) On peut le tester en testant la significativité de δ_0 dans la regression "augmentée" simplement

Ici $\hat{\delta} = 0.2163$, p-value ≈ 0. → on rejette très fortement.
↑ le coefficient de lambda

⇒ on rejette l'hypothèse de sélection exogène ici, la sélection est significativement endogène.

EM2 5.17(N) Censure et Sélection

On a une corrélation positive assez forte ($\hat{\rho} = 0.535$) entre les deux erreurs ε et η

i.e. Y^* tend à être grand quand D est égal à 1 (petit) (0)

↳ biais de sélection positif

C'est logique pour ce sens de la sélection endogène :

idée, dans toute la population, chaque individu a un salaire potentiel Y^* et de fait les salariés ($Y = Y^*$) sont ceux qui ont les salaire potentiel les plus élevés

dans l'autre sens, ceux qui ne travaillent pas sont ceux qui ont un salaire potentiel plus faible, anticipent un salaire plus faible

Modèle de sélection généralisée : application

Questions :

- Quel est l'effet marginal de l'expérience potentielle sur le salaire potentiel ? Réponse : La dérivée du log(salaire horaire) par rapport à l'expérience vaut 0.027-0.0004 * 2 *exp=0.027-0.0008 * exp. Lorsque exp=10, alors une année d'expérience supplémentaire augmente le salaire horaire de 1.9% (0.027-0.008).
- A quoi correspondent `lambda`, `rho` et `sigma` ? Réponse : `lambda`$=\delta_0 \equiv \frac{Cov(\epsilon,\eta)}{V(\eta)} = Cov(\epsilon,\eta)$; `sigma`$=\sqrt{V(\epsilon)}$; `rho`$=\frac{Cov(\epsilon,\eta)}{\sqrt{V(\epsilon)V(\eta)}} = \frac{\delta_0}{\sqrt{V(\epsilon)}}$.
- La sélection est-elle significativement endogène ici ? Dans quel sens joue-t-elle ? Réponse : Oui car on rejette l'hypothèse nulle que $\delta_0 = 0$. Comme $\widehat{delta} = 0.217$ (signe positif), le salaire horaire *potentiel* des femmes non-actives est inférieur à celui des femmes actives. toutes choses égales par ailleurs

non active η petit ↑ ε petit également ↳ salaire plus faible

- L'instrument a-t-il un effet significatif ? Réponse : Oui, la variable `NBENF6` est statistiquement significative.