Correction – Quiz 1 sur le Chapitre 1
(GMM et compléments sur les variables instrumentales)

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

ENSAE 2A – Économétrie 2 – Printemps 2022

Sauf mention contraire, les notations utilisées cherchent à reprendre celles du cours. Pour ce premier chapitre, vous pouvez également regarder les quiz 9, 10 et 11 d'Économétrie 1 sur les variables instrumentales.

Question 1 (deux conditions d'homoscédasticité)

En Économétrie 1 (EM1), au Chapitre 2, slide 6, a été introduite la condition d'homoscédasticité suivante

$$\mathbb{E}[\varepsilon^2 XX'] = \mathbb{E}[\varepsilon^2]\,\mathbb{E}[XX'] \tag{1}$$

On parle parfois pour (1) d'homoscédasticité *faible* par opposition à la condition d'homoscédasticité *forte* (voir Chapitre 1 d'Économétrie 2 (EM2), slide 8) suivante

$$\mathbb{E}[\varepsilon^2 \,|\, X] = \mathbb{E}[\varepsilon^2] \overset{\text{noté}}{=:} \sigma^2 \tag{2}$$

(a) Montrez que l'homoscédasticité forte (2) implique l'homoscédasticité faible (1).

L'idée est d'utiliser la loi des espérances itérées en conditionnant par la variable pour laquelle on peut utiliser l'hypothèse, c'est-à-dire ici en conditionnant par X pour utiliser $\mathbb{E}[\varepsilon^2 \,|\, X] = \sigma^2$. Supposons (2). On a alors

$$\begin{aligned}
\mathbb{E}[\varepsilon^2 XX'] &= \mathbb{E}[\mathbb{E}[\varepsilon^2 XX' \,|\, X]] \quad \text{(loi des espérances itérées)} \\
&= \mathbb{E}[\mathbb{E}[\varepsilon^2 \,|\, X] XX'] \quad \text{(linéarité de l'espérance conditionnelle)} \\
&= \mathbb{E}[\sigma^2 XX'] \quad \text{(hypothèse d'homoscédasticité forte)} \\
&= \sigma^2 \mathbb{E}[XX'] \quad \text{(linéarité de l'espérance, } \sigma^2 \text{ est une constante)} \\
&= \mathbb{E}[\varepsilon^2]\mathbb{E}[XX'].
\end{aligned}$$

L'homoscédasticité forte implique donc l'homoscédasticité faible, d'où ces adjectifs !

De manière similaire (exercice à faire en plus), on peut montrer que ce qu'on appelle parfois l'exogénéité *faible* du régresseur X implique l'exogénéité qualifiée parfois par opposition de *forte*.[1]

Exogénéité (parfois, "exogénéité faible") : les régresseurs X et *le terme d'erreur* ε *dans le modèle causal / structurel / la représentation linéaire où intervient le paramètre d'intérêt* β_0 *qu'on cherche à estimer* sont orthogonaux[2], ce qui est équivalent à être non corrélés dès lors que ε est centré (sans perte de généralité si le modèle contient une constante) :

$$\mathbb{E}[X\varepsilon] = 0 \overset{\text{si } \varepsilon \text{ centré}}{\iff} \mathbb{C}\text{ov}(X, \varepsilon) = 0. \tag{3}$$

1. N.B. : cette terminologie s'emploie avec des données en coupe et non des panels ; à ne pas confondre avec la terminologie utilisée dans le cadre du Chapitre 2 avec la notion d'exogénéité stricte par opposition à faible.

2. Pour le produit scalaire (au sens propre pour des variables univariées U et V de L^2 seulement mais s'étend à une variable aléatoire réelle et un vecteur aléatoire en raisonnant composante par composante) : $\langle U, V \rangle = \mathbb{E}[UV]$.

Exogénéité forte : *le terme d'erreur ε dans le modèle causal / structurel / la représentation linéaire où intervient le paramètre d'intérêt β_0 qu'on cherche à estimer* est "mean-independent" ("indépendant en moyenne, en espérance") des régresseurs X, c'est-à-dire que l'espérance conditionnelle de ε par rapport à X est simplement égale à son espérance inconditionnelle (laquelle est égale à 0, sans perte de généralité à nouveau, si le modèle contient une constante) :

$$\mathbb{E}[\varepsilon \mid X] = \mathbb{E}[\varepsilon] = 0. \tag{4}$$

(b) Comment est appelée la relation $\mathbb{E}[\varepsilon \mid X] = \mathbb{E}[\varepsilon] = 0$ entre les variables aléatoires ε et X ? Discutez ses liens avec (i) $\mathbb{C}\text{ov}(X, \varepsilon) = 0$: la non-corrélation entre ε et X (*rappel* : on peut supposer sans perte de généralité ε centré dès lors que les régresseurs contiennent une constante) et avec (ii) $\varepsilon \perp\!\!\!\perp X$: l'indépendance entre ε et X.

Cette question est simplement là pour faire quelques rappels sur le degré de (non-)dépendance entre des variables aléatoires.

Rappels (degrés de non-dépendance)

1. l'indépendance $X \perp\!\!\!\perp \varepsilon$, qui implique
2. la "mean-independence" ("indépendant en moyenne/espérance"), l'espérance conditionnelle est simplement égale à l'espérance (inconditionnelle) : $\mathbb{E}[\varepsilon \mid X] = \mathbb{E}[\varepsilon]$, qui implique
3. la non-corrélation dès lors que ε est centrée : $\mathbb{E}[X\varepsilon] = \mathbb{C}\text{ov}(X, \varepsilon) = 0$.

Les réciproques sont fausses en général. Il existe néanmoins des cas particuliers où elles sont vérifiées, voir par exemple :

— le cas des vecteurs gaussiens ;

— le cas où X est une variable binaire (voir slides de correction du TD5 d'Économétrie 1).

Question 2 (l'hypothèse d'exogénéité)

Cette question cherche à faire le lien entre le cours d'EM1 et le cours d'EM2. Elle permet de revenir sur des notions importantes.

Le cours d'EM1 distinguait deux représentations linéaires qui sont, en général, sans hypothèses supplémentaires liées à l'absence de biais de sélection, *distinctes* (voir notamment le slide 14 du Chapitre 3 d'EM1 et plus largement ce chapitre 3 du cours d'EM1) :

P. *la représentation non-causale, simple Projection linéaire* où, par construction [3] de la régression linéaire théorique (aussi appelée projection linéaire), le résidus dans cette représentation non-causale est orthogonal [4] avec les régresseurs (voir notamment la Proposition 5, slide 29 du Chapitre 1 d'EM1).

C. *la représentation Causale* qui fait intervenir le paramètre causal d'intérêt, celui qu'on cherche à estimer, et dans laquelle le terme d'erreur n'est pas automatiquement orthogonal avec le régresseur. Ce terme d'erreur est celui qui intervient dans l'équation avec les

3. On a toujours cela par construction / définition de la projection sous réserve que les conditions de moments habituelles sont vérifiées : existence d'un moment d'ordre 2 fini pour la variable expliquée, existence d'un moment d'ordre 2 fini pour les variables explicatives ou régresseurs, absence de colinéarité parfaite des régresseurs (voir l'énoncé de la Proposition 5 du Chapitre 1 d'EM1 pour l'écriture formelle de ces conditions).

4. Par la suite, on dira que deux variables aléatoires U et V sont orthogonales si $\mathbb{E}[UV] = 0$. Dès lors qu'une des deux variables est centrée (ce qui sera le cas des résidus sans perte de généralité lorsque le vecteur des régresseurs inclut une constante), on a $\mathbb{E}[UV] = \mathbb{C}\text{ov}(U, V)$ et l'orthogonalité ainsi définie est donc équivalente à la non-corrélation entre les variables U et V.

variables potentielles. Les variables potentielles et le paramètre causal sont définis conjointement. Ce terme d'erreur capte l'hétérogénéité inobservée, c'est-à-dire le fait que deux individus avec les mêmes variables explicatives auront néanmoins la plupart du temps des variables expliquées différentes.

Dans cette représentation causale, on n'a pas automatiquement l'orthogonalité du régresseur et de ce terme d'erreur. *C'est une hypothèse de supposer cela* et c'est ce qu'on appelle *l'hypothèse d'exogénéité* (qui est liée à l'hypothèse d'absence de biais de sélection).

Si on ne la fait pas, les deux représentations diffèrent et l'estimateur MCO ne permet pas d'identifier le paramètre causal d'intérêt. Si on la fait, les deux représentations coïncident et l'estimateur MCO identifie le paramètre causal d'intérêt (il converge en probabilité vers ce paramètre).

On trouve également deux représentations linéaires dans le Chapitre 1 du cours d'EM2 :

— au slide 5

$$Y = X'\beta_0 + \varepsilon \tag{5}$$

— au slide 10

$$Y = X'\tilde{\beta} + \tilde{\varepsilon} \tag{6}$$

Avec les notations utilisées dans le cours d'EM2,

(a) Laquelle est la représentation simple projection linéaire P. ?

(b) Laquelle est la représentation causale C. ?

(c) Donnez l'expression de l'estimateur MCO, $\widehat{\beta}_{\text{MCO}}$, de la régression de Y sur X.

(d) Quel paramètre, entre β_0 et $\tilde{\beta}$, est *toujours* (sous les conditions de moment habituelles et en supposant un échantillonnage i.i.d. des unités, hypothèses qui seront toujours faites dans la suite) la limite en probabilité de l'estimateur $\widehat{\beta}_{\text{MCO}}$?

(e) Avec les notations des équations (5) et (6), écrivez ce qu'on appelle "la condition d'exogénéité du régresseur X".

(f) Sous cette hypothèse d'exogénéité, que pouvez-vous dire de β_0 et de $\tilde{\beta}$?

(g) Donnez un exemple classique d'organisation de collecte des données où cette hypothèse a des chances d'être vérifiée.

(a) La représentation causale C. est le modèle du slide 5 :

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

Dans l'ensemble du chapitre, β_0 est le paramètre causal qu'on cherche à estimer. [5] On n'a donc pas nécessairement X et ε non corrélés. Supposer cela, c'est faire l'hypothèse que X est exogène, c'est-à-dire de manière sous-entendu, exogène *pour le modèle* (5)*, c'est-à-dire relativement au terme d'erreur ε qui capte l'hétérogénéité individuelle inobservée.*

(b) La représentation simple projection linéaire P. est celle du slide 10 où justement le slide insiste sur le fait qu'on peut *toujours* (sous les conditions de moment habituelles X dans L^2, Y dans L^2 et $\mathbb{E}[XX']$ inversible) définir un vecteur de coefficient $\tilde{\beta}$ et une variable aléatoire réelle $\tilde{\varepsilon}$ par [6]

— $\tilde{\beta} := \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$, et

5. Remarque : par rapport au modèle plus général avec des effets hétérogènes du Chapitre 3 d'EM1, on suppose donc un effet causal homogène.

6. Remarquez qu'on définit bien ces deux objets $\tilde{\beta}$ et $\tilde{\varepsilon}$ à partir seulement de la distribution jointe de (X, Y).

— $\tilde{\varepsilon} := Y - X'\tilde{\beta}$

tels qu'alors

$$Y = X'\tilde{\beta} + \tilde{\varepsilon} \quad \text{avec } \mathbb{E}[X\tilde{\varepsilon}] = 0.$$

(c) et (d) Si on dispose d'un échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} (X, Y)$, alors l'estimateur des MCO de Y sur X est défini par

$$\widehat{\beta}_{\text{MCO}} := \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

et il converge toujours (sous les conditions de moment usuelles supposées toujours vérifiées ici) en probabilité quand la taille de l'échantillon n tend vers $+\infty$ vers $\tilde{\beta}$.

(e) Mais en général, $\tilde{\beta} \neq \beta_0$. Pour avoir l'égalité, il faut et il suffit que les deux représentations P. et C. coïncident et pour cela il faut et il suffit que le résidus dans la représentation C. soit bien orthogonal au régresseur, c'est la condition d'exogénéité du régresseur X : $\mathbb{E}[X\varepsilon] = 0$.

(f) Sous cette hypothèse, on a $\tilde{\beta} = \beta_0$, et également $\tilde{\varepsilon} = \varepsilon$, et l'estimateur MCO, $\widehat{\beta}_{\text{MCO}}$, de Y sur X est donc bien un estimateur consistant de β_0, le paramètre causal d'intérêt.

(g) En général, il n'y a pas de raison d'avoir $\mathbb{E}[X\varepsilon] = 0$ car "tout bouge en même temps", les variables explicatives peuvent être liées, corrélées avec les facteurs inobservables. Les individus avec des ε élevés peuvent par exemple avoir aussi tendance, car ils choisissent d'une manière ou d'une autre X mettons, à avoir des X faibles.

Exemple d'une telle situation (voir aussi les TD d'EM1 sur ce thème, données de LaLonde) : Y est le salaire horaire, donc ε élevé signifie que cet individu a, pour des raisons individuelles spécifiques non observées par l'économètre, un salaire élevé, et X est binaire, égal à 1 si l'individu participe à un programme de formation pour l'aider sur le marché du travail, 0 sinon. Si on laisse les gens choisir, on peut penser que les personnes qui vont décider de suivre ce programme de formation $X = 1$ sont plutôt ceux avec des salaires faibles (pour tenter d'augmenter leur salaire). Mais, on pourrait aussi penser et défendre le contraire ! Et c'est justement le problème : on ne sait pas dans un tel cas avec des données non-expérimentales. Il se pourrait que ce soit les individus avec des ε élevés, ce qui capterait le fait qu'ils soient motivés/travailleurs/etc. qui sont les plus à même de suivre la formation et de la valider et donc d'avoir $X = 1$ (X élevé).

Un exemple classique d'organisation de collecte des données où l'hypothèse d'exogénéité a des chances d'être satisfaite est le cas des expériences aléatoires (voir aussi la partie du Chapitre 3 d'EM1 sur ce thème). En effet, on choisit alors le régresseur X aléatoirement et, si la procédure est bien réalisée, on peut alors supposer que X est bien déterminé aléatoirement, indépendamment de toute autre variable et en particulier du résidus ε.

Question 3 (discret ou qualitatif, continu ou quantitatif)

La slide 3 du Chapitre 1 d'EM2 est importante car elle présente les notations qui seront utilisées tout au long du cours, y compris dans les chapitres suivants.

Elle dit notamment : "Les composantes de X (le vecteur colonne des régresseurs) pourront être continues ou discrètes. Dans le deuxième cas, on inclura toutes les indicatrices des modalités correspondantes, sauf une. Exemple : pour la variable d'activité (position sur le marché du travail selon le sens du Bureau International du Travail) prenant trois valeurs (actif en emploi, actif au chômage, inactif), on inclura, par exemple (au sens où on aurait pu choisir autrement les deux modalités incluses dans les régresseurs) $\mathbb{1}_{\text{en emploi}}$ et $\mathbb{1}_{\text{au chômage}}$."

Ainsi, si on dit qu'on fait la régression d'une certaine variable expliquée Y sur la position sur le marché du travail, cela signifie qu'on fait formellement, par exemple, si on choisit la modalité "inactif" comme modalité de référence (au sens où elle correspondra à la constante), la régression de Y sur $X = (1, X_{\text{employé}}, X_{\text{chômeur}})'$ où, pour tout individu i, $X_{\text{employé},i} := \mathbb{1}\{i \text{ est en emploi}\}$ et $X_{\text{chômeur},i} := \mathbb{1}\{i \text{ est au chômage}\}$.

Imaginons (il s'agit ici de résultats entièrement inventés) que Y soit le nombre de minutes quotidiennes de lecture (de livres), en moyenne sur une semaine de référence à laquelle on réalise l'enquête, et qu'on obtient comme coefficients estimés par MCO pour cette régression :

$$\widehat{Y} = 42 - 7X_{\text{employé}} + 3X_{\text{chômeur}} \tag{7}$$

ce qui est une façon compacte d'écrire que les réalisations des estimateurs MCO des coefficients associés à 1 (la constante), $X_{\text{employé}}$ et $X_{\text{chômeur}}$ sont respectivement $+42$, -7 et $+3$.

(a) Donnez une interprétation quantitative, au moyen d'une phrase syntaxiquement correcte composée de mots et sans symboles mathématiques, de ces trois coefficients estimés.

La constante correspond ici à la modalité de référence, lorsque $X_{\text{employé}}$ et $X_{\text{chômeur}}$ sont tous les deux nuls, c'est-à-dire pour un individu inactif sur le marché du travail.

Sur l'échantillon étudié, la moyenne empirique du temps de lecture des inactifs est de 42 minutes quotidiennes. En allant au-delà de l'échantillon et pour une interprétation sur la population d'intérêt, on peut interpréter ce coefficient ainsi : au sein de la population étudiée (celle du champs de l'enquête, du sondage par lequel on a collecté les données), on estime le temps moyen de lecture pour les individus inactifs à 42 minutes quotidiennes.

Avec des variables qualitatives, il faut prendre garde à interpréter les coefficients relativement à la modalité de référence (celle pour laquelle on n'inclut pas d'indicatrices parmi les régresseurs).

Ainsi, on interprète le -7 comme : toutes choses égales par ailleurs, on prédit un temps de lecture quotidien inférieur de 7 minutes pour un individu employé *par rapport à un individu inactif*. D'après ce qui précède, on peut donc aussi dire : au sein de la population étudiée, on estime à $35 = 42 - 7$ minutes le temps moyen de lecture des individus actifs employés.

De même, pour le coefficient $+3$ estimé devant $X_{\text{chômeur}}$: toutes choses égales par ailleurs, *par rapport à un individu inactif*, on prédit un temps de lecture quotidien supérieur de 3 minutes pour un individu chômeur.

On peut aussi remarquer qu'on prédit un temps de lecture supérieur de $10 = 3 - (-7)$ minutes pour un individu chômeur par rapport à un individu employé.

Aussi, la moyenne parmi les individus actif chômeurs du temps de lecture quotidien est estimée à $45 = 42 + 3$ minutes.

(b) Pour quelle raison exclut-on une des modalités de la variable "position sur le marché du travail" dans la régression précédente ?

Si on incluait les indicatrices de chacune des modalités, on ferait la régression de Y sur $X = (1, X_{\text{inactif}}, X_{\text{employé}}, X_{\text{chômeur}})'$. Or, une telle régression n'est pas bien définie uniquement car la condition d'absence de colinéarité parfaite entre les régresseurs n'est pas satisfaite (on parle en anglais dans ce genre de situations de "*dummy variable trap*"). En effet, puisqu'il n'y a que trois possibilités mutuellement exclusives sur le marché du travail (dans la classification retenue de l'activité au sens du Bureau International du Travail), on a

$$1 = X_{\text{inactif}} + X_{\text{employé}} + X_{\text{chômeur}},$$

soit encore

$$X'\lambda = 0 \text{ pour } \lambda = (-1, 1, 1, 1)' \neq 0_{\mathbb{R}^4}.$$

C'est pourquoi, dès lors qu'on choisit d'inclure une constante dans le modèle, il faut exclure une des modalités lorsqu'on inclut une variable qualitative comme variable explicative dans une régression linéaire.

(c) Notons X_{inactif} l'indicatrice qu'un individu soit inactif. Est-il possible de faire la régression linéaire de Y sur $(X_{\text{inactif}}, X_{\text{employé}}, X_{\text{chômeur}})'$ ou cela viole-t-il une des conditions de moment requises ? Le cas échéant, laquelle ? Sinon, si cela est possible à partir de la régression précédente (7), indiquez ce qu'on aurait obtenu comme estimation par MCO des coefficients dans cette régression.

Oui, on peut faire la régression linéaire de Y sur $X = (X_{\text{inactif}}, X_{\text{employé}}, X_{\text{chômeur}})'$. Il s'agit d'une régression sans constante, et il faut alors inclure les indicatrices pour chacune des modalités. Ce serait une autre manière de procéder. Dans le cours, on ne la suivra pas et on fera la précédente régression : avec une constante et en excluant une des modalités (qui joue alors le rôle de modalité de référence).

On peut retrouver directement les coefficients estimés de la régression de Y sur $X = (X_{\text{inactif}}, X_{\text{employé}}, X_{\text{chômeur}})'$ à partir de (7). Il s'agit en fait d'une simple reparamétrisation, qu'on peut voir en partant de (7) et en utilisant $1 = X_{\text{inactif}} + X_{\text{employé}} + X_{\text{chômeur}}$:

$$\begin{aligned}
\widehat{Y} &= 42 - 7X_{\text{employé}} + 3X_{\text{chômeur}} \\
&= 42 \times 1 - 7X_{\text{employé}} + 3X_{\text{chômeur}} \\
&= 42 \times (X_{\text{inactif}} + X_{\text{employé}} + X_{\text{chômeur}}) - 7X_{\text{employé}} + 3X_{\text{chômeur}} \\
\widehat{Y} &= 42X_{\text{inactif}} + 35X_{\text{employé}} + 45X_{\text{chômeur}}
\end{aligned}$$

On aurait ainsi obtenu un coefficient estimé de 42 pour X_{inactif}, 35 pour $X_{\text{employé}}$ et 45 pour $X_{\text{chômeur}}$. Ces estimés sont les moyennes empiriques dans l'échantillon des temps lecture pour chacun des trois sous-groupes.

Le cours parle de variable *continue* et *discrète*. Toutefois, peut-être est-il préférable d'entendre ces mots davantage dans un sens ordinaire, en pensant à l'opposition variable *quantitative* (pour continue) contre variable *qualitative* (pour discrète), plutôt que dans un sens mathématique formel (admettre une densité par rapport à la mesure de Lebesgue ou bien par rapport à la mesure de comptage).

(d) Pour certaines variables entièrement qualitatives (comme l'exemple précédent de la position sur le marché du travail), il est indispensable de les inclure au moyen des indicatrices de chacune des modalités sauf une. Se convaincre que ne pas faire cela serait absurde ; vous pouvez penser à la façon en pratique de coder ces variables dans une base de données.

La position du marché du travail est une variable qualitative, n'ayant pas de sens quantitatif. Dans une base de données, on pourra néanmoins fréquemment l'encoder au moyen de nombres, disons par une variable P (pour Position) égale, par exemple à 1 si l'individu est inactif, 2 si actif occupé et 3 si actif chômeur.

Toutefois, ces chiffres n'ont aucun sens en tant que tels. Si on faisait à tort une régression de Y sur P codée ainsi, on supposerait en faisant cela :

— que l'effet (en termes de modification de prédiction) de passer de $P = 1$ à $P = 2$ (inactif à actif occupé) est le même (effet linéaire) que de passer de $P = 2$ à $P = 3$ (actif occupé à chômeur) ;

— que l'effet de passer, par exemple, d'inactif à actif occupé est deux fois moindre que de passer d'inactif ($P = 1$) à actif chômeur ($P = 3$).

Pour d'autres variables ayant un sens quantitatif mais gardant un aspect un peu discret (disons, au sens de prendre un nombre relativement limité de valeurs distinctes dans les données), il est intéressant de comprendre la différence de modélisation et d'interprétation entre

(i) les inclure directement en tant que telle, en tant que nombre, en les voyant comme des variable continues / quantitatives, et

(ii) inclure les indicatrices de chacune de leurs modalités sauf une, en les voyant comme des variables discrètes / qualitatives.

(e) Gardons Y le temps moyen de lecture quotidien comme variable expliquée (ou vous pouvez en choisir une autre). On s'intéresse à l'effet (effet dans un sens prédictif du moins) du nombre d'enfants sur le temps moyen de lecture. Imaginons, par exemple, qu'on se restreint à une population d'individus ayant entre 0 et 3 enfants. Expliquez la différence de modélisation et d'interprétation entre faire les régressions

(i) de Y sur une constante et la variable N où N_i est le nombre d'enfants de l'individu i (le support de la variable N est donc $\{0, 1, 2, 3\}$) ;

(ii) de Y sur une constante et les variables X_1, X_2 et X_3 où, pour $k \in \{1, 2, 3\}$, on pose $X_{k,i} = \mathbb{1}\{\text{l'individu } i \text{ a } k \text{ enfant(s)}\}$.

Y a-t-il une modélisation plus restrictive qu'une autre ? En d'autres termes, est-ce qu'une des deux modélisations fait (implicitement) une hypothèse quant à l'effet du nombre d'enfants sur le temps de lecture ?
Indice : relisez et repensez aux slides 18 et 19 du Chapitre 3 d'EM1 ; que signifie un effet linéaire ?

Le nombre d'enfants est une variable ayant un sens quantitatif : les chiffres 0, 1, 2 et 3 ont un sens quantitatif (bien voir la différence par rapport à la simple convention d'encodage pour la variable de position sur le marché du travail P de l'exemple précédent). C'est toutefois une variable qu'on peut qualifiée de discrète au sens où elles prend des valeurs dans un ensemble fini et qui est petit relativement au nombre de données (par opposition, le salaire horaire sera une variable quantitative continue et non discrète typiquement car on observera la plupart du temps beaucoup de valeurs différentes du salaire horaire relativement au nombre d'observations). Dans ce cas, on peut envisager les deux modélisations (i) et (ii).

Modélisation (i) : inclure la variable en tant que telle comme une variable *quantitative*. Par définition d'un effet linéaire, cette modélisation (dans un modèle level-level ici – voir aussi les extensions permises par des modèles log-level, log-log, etc.) impose que l'effet d'avoir un enfant supplémentaire sur la variable de résultat Y est le même qu'on passe de 0 à 1, 1 à 2, ou 2 à 3 enfants. Par exemple, imaginons qu'on obtienne en estimant par MCO ce modèle

$$\widehat{Y} = 44 - 8N.$$

On prédira alors, toutes choses égales par ailleurs, une variation du temps de lecture de -8 minutes pour chaque enfant supplémentaire.

Modélisation (ii) : inclure la variable en tant que variable *qualitative* en incluant les indicatrices de chacune de ses modalités, sauf une. En comparaison, ce modèle est plus flexible : il y a un paramètre associé à chaque modalité permettant ainsi un effet spécifique à chaque modalité. On n'impose donc pas un effet linéaire ; on autorise un effet différent sur la variable expliquée Y de passer de 0 à 1 enfants que de passer de 2 à 3 – on peut imaginer que le fait de devenir parent a un effet différent, induit un changement de comportement plus important que de passer de 2 à 3 enfants où les parents commencent à avoir l'habitude et l'enfant supplémentaire provoque une modification plus marginale de leurs comportements. On pourrait ainsi estimer par exemple (chiffres inventés ici) pour ce modèle :

$$\widehat{Y} = 44 - 13X_1 - 19X_2 - 17X_3.$$

Par rapport à ne pas avoir d'enfant, on prédit une diminution de 13 minutes du temps de lecture en ayant un enfant. Toutes choses égales par ailleurs, si un individu a 2 enfants au lieu de 1, on prédit une variation du temps de lecture de $-19-(-13)=-6$ minutes, soit une diminution de 6 minutes du temps de lecture quotidien.

La modélisation (i) est une modélisation plus restrictive que la modelisation (ii).

Dans Stata, si `Y` et `N` sont les noms des variables Y et N dans la base de données, la modélisation (i) est implémentée par la commande

```
regress Y N, robust
```

alors qu'on peut implémenter la modélisation (ii) directement (sans avoir à créer manuellement les indicatrices) grâce à l'opérateur `i.` ainsi (nous reverrons des exemples en TD) :

```
xi : regress Y i.N, robust
```

(f) (bonus) Retrouvez les formes originales et auteurs de ces deux citations [7] :

"*Le quantitatif est un qualitatif impur*" – Paul Valéry, "le fond est une forme impure" (voir aussi la conception de la littérature comme forme-sens).

"*Le qualitatif, c'est le quantitatif qui remonte à la surface*" – Victor Hugo, "La forme, c'est le fond qui remonte à la surface".

Question 4 (MCO, 2MC, GMM et efficacité asymptotique)

On considère le modèle linéaire du cours

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

où X est un vecteur aléatoire de dimension K.

On ne suppose pas l'exogénéité de X : $\mathbb{E}[X\varepsilon] \neq 0$ a priori. On ne suppose pas non plus l'homoscédasticité des résidus ε. Par contre, on suppose qu'on dispose d'un instrument Z, vecteur aléatoire de dimension $L \geq K$, qui est valide : $\mathbb{E}[ZX']$ est de plein rang et $\mathbb{E}[Z\varepsilon] = 0$.

Ici, X est endogène : on ne suppose pas l'exogénéité de X, on a $\mathbb{E}[X\varepsilon] \neq 0$. L'estimateur MCO de Y sur X n'estime donc pas β_0 de manière convergente. Pour la même raison, la condition de moment $\mathbb{E}[X(Y - X'\beta_0)] = 0$ n'est pas valide ici, justement on n'a pas $\mathbb{E}[X\varepsilon] = 0$.

Un estimateur *asymptotiquement efficace / optimal* (les deux adjectifs sont utilisés comme synonymes ici) de β_0 est un estimateur consistant de β_0, asymptotiquement normal autour de β_0, et avec la variance asymptotique la plus faible possible (au sens des matrices semi-définies positives) parmi la classe des estimateurs asymptotiquement normaux (grosso modo). Il doit donc en particulier être consistant. Or, l'estimateur GMM utilisant la condition de moment $\mathbb{E}[X(Y - X'\beta_0) = 0]$ est exactement l'estimateur MCO, il n'est donc pas convergent ici, X n'étant pas exogène.

On dispose par contre d'un instrument Z valide (condition de rang et exogénéité). L'estimateur des 2MC obtenu en instrumentant X par Z estime donc bien de manière consistante β_0 : c'est un estimateur consistant de β_0.

Se pose par contre la question supplémentaire de son efficacité/optimalité asymptotique : l'estimateur 2MC a-t-il ici la plus faible variance asymptotique ? La réponse se trouve aux slides 30 et 31 du Chapitre 1 d'EM2.

7. Vous pouvez inverser "quantitatif" et "qualitatif" si vous trouvez cela mieux.

Si les résidus sont homoscédastiques (relativement à l'instrument Z, c'est-à-dire $\mathbb{E}[\varepsilon^2 ZZ'] = \mathbb{E}[\varepsilon^2]\mathbb{E}[ZZ']$), l'estimateur 2MC est asymptotiquement efficace/optimal.

Sinon, il ne l'est pas et l'estimateur asymptotiquement optimal s'obtient par la méthode des GMM en deux étapes en utilisant la condition de moment $\mathbb{E}[Z(Y - X'\beta_0)] = 0$, qui est bien la condition d'exogénéité de Z : $\mathbb{E}[Z\varepsilon] = 0$.

(a) Alors

1. L'estimateur des moindres carrés ordinaires de Y sur X estime β_0 de manière convergente – **Faux**, X n'est pas supposé exogène.
2. L'estimateur des doubles moindres carrés obtenu en utilisant Z comme instrument de X estime β_0 de façon convergente et est asymptotiquement optimal – **Faux**, il est consistant (instrument Z supposé valide) mais il n'est pas asymptotiquement optimal car les résidus ne sont pas supposés homoscédastiques.
3. Un estimateur asymptotiquement optimal de β_0 peut être obtenu par la méthode des moments généralisés en utilisant la condition $\mathbb{E}[X(Y - X'\beta_0)] = 0$ – **Faux**, X n'est pas exogène et cette condition de moment n'est pas satisfaite par β_0 et ne peut donc pas être utilisée pour construire un estimateur GMM consistant.
4. Un estimateur asymptotiquement optimal de β_0 peut être obtenu par la méthode des moments généralisés en utilisant la condition $\mathbb{E}[Z(Y - X'\beta_0)] = 0$ – **Vrai**, il s'obtient en deux étapes (estimateur GMM optimal en deux étapes, voir notamment les slides 28, 30 et 31 du Chapitre 1).

(b) Que signifie un estimateur *asymptotiquement optimal* de β_0 ? Voir ci-dessus ; voir aussi votre cours de Statistique 1 du premier semestre à propos de l'estimateur du maximum de vraisemblance dans des modèles réguliers.

Question 5 (un exemple pratique d'estimateur 2MC)

On considère le modèle linéaire

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

avec X_1, X_2, Z_1 et ε des variables aléatoires réelles vérifiant $\mathbb{E}[X_1\varepsilon] \neq 0$, $\mathbb{E}[X_2\varepsilon] = 0$ et $\mathbb{E}[Z_1\varepsilon] = 0$. On cherche à estimer les paramètres d'intérêt β_0, β_1 et β_2.

La façon pratique de faire les 2MC est bien détaillée dans le cours, notamment avec l'exemple aux slides 15 et 16 du Chapitre 1. Ici, dans le contexte de cette question,

— les régresseurs endogènes sont : X_1,

— les régresseurs exogènes sont : X_2,

— les instruments à proprement parler sont : Z_1.

Avec les notations du cours, on a donc $X = (1, X_1, X_2)'$ et $Z = (1, Z_1, X_2)'$ (également 1 car on met toujours une constante, on peut la voir comme un des régresseurs exogènes formellement). *Attention* : la notation Z diffère entre les cours d'Économétrie 2 et d'Économétrie 1. Dans votre cours d'Économétrie 2, la notation Z correspond à *un vecteur colonne qui regroupe à la fois les variables de contrôle supposées exogènes et les instruments en tant que tel* ; d'où $Z = (1, Z_1, X_2)'$.

En première étape, on régresse *chacun* des régresseurs endogènes sur Z, c'est-à-dire sur toutes les variables exogènes : "régresseurs exogènes" et "instruments à proprement parler".

En deuxième étape, on régresse la variable expliquée Y sur les prédicteurs obtenus en première étape des régresseurs endogènes *et* sur les "régresseurs exogènes". Attention à ne pas

oublier d'inclure également les régresseurs supposés exogènes dans la régression de deuxième étape.

(a) Sous ces hypothèses,

1. l'estimateur 2MC s'obtient en régressant X_1 sur X_2 et Z_1, puis en régressant Y sur la valeur prédite obtenue $\widehat{X}_1$ et X_2 – **Vrai**.
2. l'estimateur 2MC s'obtient en régressant X_1 sur Z_1, puis en régressant Y sur la valeur prédite obtenue $\widehat{X}_1$ et X_2 – **Faux**, il manque dans cette réponse le régresseur exogène X_2 dans la régression de première étape.
3. l'estimateur 2MC s'obtient en régressant X_1 sur X_2 et Z_1, puis en régressant Y sur la valeur prédite obtenue $\widehat{X}_1$ – **Faux**, il manque dans cette réponse le régresseur exogène X_2 dans la régression de deuxième étape.
4. Puisque X_2 est supposée exogène, on peut l'utiliser comme instrument de X_1 sans avoir recours à Z_1 et obtenir ainsi des estimateurs convergents de β_0, β_1 et β_2 – **Faux**, X_2 ne peut pas être un instrument valide en général car il influence Y directement (dès lors que $\beta_2 \neq 0$) et non uniquement via X_1. De plus, plus fondamentalement, on cherche ici à estimer $(\beta_0, \beta_1, \beta_2)$, et en particulier donc β_1 qu'on peut interpréter comme l'effet de X_1 sur Y en contrôlant par X_2, à X_2 fixé et non l'effet de X_1 sur Y sans contrôler par X_2, qui serait un autre paramètre d'intérêt.
5. Puisqu'on a deux variables explicatives X_1 et X_2 mais un seul instrument Z_1, on ne peut pas construire l'estimateur 2MC de $(\beta_0, \beta_1, \beta_2)$ – **Faux**, il faut au moins autant d'instruments à proprement parler que de variables explicatives ou régresseurs *endogènes*. Ici, on cherche à estimer un modèle avec deux variables explicatives X_1 et X_2, mais X_2 est supposé être exogène et on a donc un seul régresseur endogène X_1 ; il faut donc au moins un instrument à proprement parler pour X_1, qui est bien le Z_1 proposé ici et qui permet de construire l'estimateur 2MC de $(\beta_0, \beta_1, \beta_2)$ en instrumentant X_1 par Z_1

(b) Écrivez la commande Stata correspondant à l'estimateur 2MC de $(\beta_0, \beta_1, \beta_2)$ en prenant garde à préciser une option (au sens de la syntaxe Stata) qui permettra de vérifier une des conditions que doit satisfaire Z_1 pour être un instrument valide. Expliquez alors comment voir le résultat de ce test dans la sortie Stata obtenue.

On note respectivement `y`, `x1`, `x2` et `z1` les noms des variables Y, X_1, X_2 et Z_1 dans notre base de données Stata. On utilise la commande Stata `ivregress 2sls` avec les options

— `robust` : pour calculer des erreurs-types robustes à l'hétéroscédasticité [8], sans cette option, par défaut Stata fait l'inférence en supposant l'homoscédasticité des résidus ;

— `first` : pour montrer la régression de première étape, ce qui permet de voir directement, dans le cas où il y a un seul instrument à proprement parler, le résultat du test de la condition de pertinence de l'instrument (s'il y a plusieurs instruments, il faut faire un test de Fisher et on ne peut donc pas lire directement sur la sortie Stata le résultat ; on peut faire manuellement avec `regress` la régression de première étape puis utiliser la commande `test`).

La syntaxe, pour le cas d'un seul régresseur endogène, est la suivante

```
ivregress 2sls y (x1 = z1) x2, robust first
ivregress 2sls <variable expliquée> (régresseur endogène =
    <instruments à proprement parler>) régresseurs exogènes, robust first
```

8. Remarque de détails de code mais quand même important à bien faire en pratique : il faut préciser `ivregress 2sls` lorsqu'on utilise l'option `robust` pour que cette option soit prise en compte à la fois dans la régression de première étape et de deuxième étape. Si on écrit `ivreg y (x1 = z1) x2, robust first`, alors la deuxième étape calcule des erreurs-types robustes à l'hétéroscédasticité mais non la première étape, attention.

La slide 35 du Chapitre 1 du cours détaille la réalisation du test de pertinence du ou des instruments à proprement parler dans le cas d'un seul régresseur endogène (voir également la fin du Chapitre 4 d'Économétrie 1 sur ce thème et les quiz associés).

Ici, on a un seul instrument à proprement parler et il suffit donc de voir si le coefficient associé à cet instrument Z_1 dans la régression de première étape est statistiquement significatif au moyen d'un test de Student de $H_0 : \alpha_{02} = 0$ (non-pertinence de l'instrument) contre $H_1 : \alpha_{02} \neq 0$ (pertinence de l'instrument) – avec les notations du slide 35. On souhaite donc rejeter au niveau le plus faible possible l'hypothèse nulle H_0 pour avoir un instrument pertinent. Cela se lit directement dans le tableau de sortie Stata de la régression de première étape, `First-stage regressions`, affiché lorsqu'on précise l'option `first`.

La Figure 1 montre un exemple de telle sortie en utilisant les données de l'exercice de TD d'Économétrie 1 sur les enchères de begonias et en se plaçant dans le cadre de la question :

— la variable expliquée est `lprice` (logarithme du prix de vente),

— un régresseur endogène : `lbidders` (logarithme du nombre d'enchérisseurs présents),

— un régresseur exogène : `type1` (une variable indiquant la qualité de la fleur),

— un instrument à proprement parler : `time` (le temps écoulé, mesuré en secondes depuis 6h30, lorsque a lieu la vente aux enchères).

FIGURE 1 – Exemple de sortie de la commande `ivregress 2sls` avec les options `robust` et `first`.

```
. ivregress 2sls lprice (lbidders = time) type1, robust first

First-stage regressions
-----------------------

                                                Number of obs   =         79
                                                F(   2,     76) =       4.42
                                                Prob > F        =     0.0153
                                                R-squared       =     0.1017
                                                Adj R-squared   =     0.0780
                                                Root MSE        =     0.2027

------------------------------------------------------------------------------
             |               Robust
    lbidders |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       type1 |   .0975261   .0478454     2.04   0.045     .0022337    .1928185
        time |   .0004068   .0001744     2.33   0.022     .0000596    .0007541
       _cons |   3.714357   .0881288    42.15   0.000     3.538834    3.889881
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression      Number of obs =      79
                                              Wald chi2(2)  =    1.36
                                              Prob > chi2   =  0.5058
                                              R-squared     =  0.2442
                                              Root MSE      =  .56646

------------------------------------------------------------------------------
             |               Robust
      lprice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lbidders |   1.511987   1.328477     1.14   0.255    -1.091781    4.115755
       type1 |  -.1112258   .1795542    -0.62   0.536    -.4631457     .240694
       _cons |  -.9420072   5.147034    -0.18   0.855    -11.03001    9.145994
------------------------------------------------------------------------------
Instrumented:  lbidders
Instruments:   type1 time
```

On peut lire la p-valeur du test de Student évoquée ci-dessus dans le carré rouge dans la sortie de la première étape : $0.022 = 2.2\% < 5\%$, on rejette ici à 5% l'hypothèse nulle de non-pertinence de l'instrument ; on ne la rejette toutefois pas à 1%. On est donc relativement confiant sur la condition de pertinence même si la corrélation entre variable endogène et instrument est plutôt faible (voir aussi la question des instruments faibles). Dans le carré bleu, colonne `Coef.` de la régression de deuxième étape, on lit l'estimateur 2MC de $(\beta_0, \beta_1, \beta_2)$ obtenu ici.

Question 6 (2MC et tests de conditions)

On considère le modèle linéaire

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

avec X_1 et ε des variables aléatoires réelles.

On suppose $\mathbb{E}[\varepsilon] = 0$, ce qui est sans perte de généralité puisque le modèle contient une constante β_0, mais on ne suppose pas $\mathbb{E}[X_1\varepsilon] = 0$: a priori, X_1 est donc endogène.

On dispose potentiellement d'un instrument Z_1, où Z_1 est une variable aléatoire réelle.

L'hypothèse H_0 est l'hypothèse d'exogénéité de l'instrument Z_1 : $\mathbb{C}\text{ov}(Z_1, \varepsilon) = \mathbb{E}[Z_1\varepsilon] = 0$ (ε est centré). On ne peut tester (en partie) la condition d'exogénéité que si l'on dispose de *strictement* plus d'instruments à proprement parler que de régresseurs endogènes, c'est-à-dire que si $L > K$ en reprenant les notations du Chapitre 1. Dans ce cas, on peut réaliser un test de sur-identification du modèle permettant de tester si le terme d'erreur est bien non corrélé (orthogonal) avec les instruments (voir slides 31-33 du Chapitre 1).

Ici, on a $L = K = 2$ (2 et non 1 car il y a toujours également la constante) : on a un seul instrument à proprement parler pour un seul régresseur et on ne peut donc pas réaliser de test sur-identification et tester H_0 contre H_1.

L'autre test concerne la condition de pertinence de l'instrument Z_1. Ici, dans ce cadre avec un seul régresseur endogène, un seul instrument à proprement parler et pas de régresseurs/contrôles exogènes, la condition de pertinence de l'instrument Z_1 pour instrument X_1 est $\mathbb{C}\text{ov}(Z_1, X_1) \neq 0$, soit l'hypothèse alternative H_1'.

Contrairement à l'hypothèse d'exogénéité, on peut *toujours*, et il faut le faire !, tester la condition de rang. Ici, cela revient donc à faire le test de H_0' contre H_1' au moyen d'un test de Student de significativité statistique dans la régression de première étape de X_1 sur Z_1 (voir aussi Économétrie 1, Chapitre 4 et Économétrie 2, slide 35 du Chapitre 1 pour les autres cas).

(a) Alors, afin d'estimer (β_0, β_1) de manière consistante par la méthode des Doubles Moindres Carrés (2MC) en instrumentant X_1 par Z_1

1. On peut et il faut tester l'hypothèse nulle H_0 : $\mathbb{C}\text{ov}(Z_1, \varepsilon) = 0$ contre l'alternative H_1 : $\mathbb{C}\text{ov}(Z_1, \varepsilon) \neq 0$ – **Faux**, on a ici $L = K$, on ne peut pas faire de test de sur-identification.
2. On peut et il faut tester l'hypothèse nulle H_0' : $\mathbb{C}\text{ov}(Z_1, X_1) = 0$ contre l'alternative H_1' : $\mathbb{C}\text{ov}(Z_1, X_1) \neq 0$ – **Vrai**, on peut et on doit toujours tester la condition de pertinence de l'instrument.
3. On peut tester H_0 contre H_1 et on peut tester H_0' contre H_1' mais il est inutile de faire de tels tests pour réaliser une estimation par 2MC – **Faux**, doublement même au sens où on ne peut pas tester H_0 contre H_1 ici et qu'il serait pertinent de faire de tels tests.
4. On ne peut pas tester H_0 contre H_1 et on ne peut pas non plus tester H_0' contre H_1' – **Faux**, on peut tester H_0' contre H_1'.

On se demande maintenant s'il était bien nécessaire de faire une estimation par 2MC et si l'estimateur des Moindres Carrés Ordinaires (MCO) de la régression de Y sur X_1 (et une constante toujours sous-entendu) ne permettait pas directement d'estimer de façon consistante les paramètres d'intérêt (β_0, β_1).

La suite de cette question renvoie également au cours d'Économétrie 1, la fin du Chapitre 4, slides 44 à 47. Elle cherche principalement à mettre en garde contre la confusion entre

- Le test d'exogénéité de l'*instrument* Z_1. Cela se fait au moyen d'un test de sur-identification (test de Sargan si homoscédasticité) et nécessite d'avoir strictement plus d'instruments à proprement parler que de régresseurs endogènes : $L > K$ (voir EM2, Chapitre 1, slides 32 et 33).
- Le test d'exogénéité du *régresseur/traitement* X_1 dont on n'est pas sûr qu'il soit exogène. Cela ne nécessite pas d'avoir strictement plus d'instruments à proprement parler que de régresseurs endogènes : il suffit d'avoir $L \geq K$, et se fait au moyen de la régression dite augmentée (voir EM1, Chapitre 4, slides 44-46). On est intéressé par un tel test car, si X_1 était en fait bien exogène, il serait plus efficace (en terme de variance asymptotique) de faire les MCO de Y sur X_1 directement pour estimer β_1 plutôt que de faire les 2MC en instrument X_1 par Z_1 (voir notamment slide 44 du Chapitre 4 d'Économétrie 1).

(b) On suppose désormais que Z_1 est bien un instrument valide pour X_1, écrivez ce que cela signifie (deux conditions).

L'exogénéité de l'instrument signifie que l'instrument n'est pas corrélé (orthogonal de manière équivalente s'il y a une constante et que ε est centré) avec le terme d'erreur (ici ε) intervenant dans le modèle linéaire où se trouve les paramètres d'intérêt qu'on cherche à estimer (ici (β_0, β_1)) : $\mathbb{E}[Z_1\varepsilon] = \mathbb{C}\text{ov}(Z_1, \varepsilon) = 0$.

La pertinence de l'instrument ou condition de rang s'écrit de façon générique avec les notations du cours, ici $X = (1, X_1)'$ et $Z = (1, Z_1)'$, $\mathbb{E}[ZX']$ est de plein rang, c'est-à-dire de rang K ($K = 2$ ici).

Cette expression un peu abstraite prend *différentes* (*attention*) formes pratiques plus concrètes selon les situations (voir notamment slide 35 du Chapitre 1 d'EM2 ou la question 4 du Quiz 10 d'EM1). Ici, avec un seul régresseur endogène X_1, un seul instrument à proprement parler Z_1 et pas d'autres régresseurs exogènes, $\mathbb{E}[ZX']$ de plein rang s'écrit de manière équivalente [9] comme $\mathbb{C}\text{ov}(Z_1, X_1) \neq 0$.

(c) Sous l'hypothèse que Z_1 est un instrument valide pour X_1, est-il possible de réaliser un test indiquant s'il est préférable d'estimer (β_0, β_1) par MCO ou par 2MC ? Le cas échéant, décrivez comment réaliser un tel test.

Cette question renvoie directement aux slides 44 à 46 du Chapitre 4 d'Économétrie 1. Le terme de préférable est à entendre ici comme une estimation plus précise au sens où la variance asymptotique de l'estimateur est plus petite (au sens des matrices semi-définies positives).

On suppose ici que Z_1 est bien valide (pertinent et exogène).

Si X_1 est effectivement endogène, l'estimateur MCO n'est pas consistant et il faut donc utiliser l'estimateur 2MC pour avoir une estimation consistante de (β_0, β_1).

Par contre, si X_1 est en réalité exogène, l'estimateur MCO est consistant. L'estimateur 2MC reste également consistant. Mais la variance asymptotique de l'estimateur MCO est inférieure à la variance asymptotique de l'estimateur 2MC et il est donc préférable d'utiliser les MCO.

Sous l'hypothèse que Z_1 est bien valide, on peut faire un tel test au moyen de la régression augmentée. On commence (c'est la régression de première étape) par régresser X_1 sur Z_1 et par récupérer le résidus estimé qu'on note $\widehat{\nu}$ (pour suivre les notations d'EM1 Chapitre 4). Puis,

9. Voir correction manuscrite du TD2 d'EM2 pour le détail du calcul – écrit en petit sur la première page.

on fait la régression de Y sur X_1 *et* $\widehat{\nu}$ (d'où ce terme de régression *augmentée*). Appelons ρ_0 le coefficient théorique associé à $\widehat{\nu}$ dans cette régression. On peut montrer que (voir slide 45 d'EM1, Chapitre 4, en adaptant les notations au cadre de la question), que $\rho_0 = 0 \iff \mathbb{C}\text{ov}(X_1, \varepsilon) = 0$, c'est-à-dire que l'exogénéité de X_1 est équivalente à la nullité de ρ_0. On peut donc tester $H_0 : \mathbb{C}\text{ov}(X_1, \varepsilon) = 0$ contre $H_1 : \mathbb{C}\text{ov}(X_1, \varepsilon) \neq 0$ par un test de Student de significativité statistique du coefficient ρ_0 associé au résidus estimé dans la régression augmentée.

On suppose encore que Z_1 est un instrument valide pour X_1 mais également qu'on dispose d'un deuxième instrument valide Z_2, où Z_2 est une variable aléatoire réelle.

(d) Sous ces hypothèses, reprenez votre réponse à la question (a) – en adaptant de manière appropriée les hypothèses H_0, H_1, H_0' et H_1' pour inclure également Z_2, on souhaite désormais estimer (β_0, β_1) en instrument X_1 par Z_1 et Z_2.
Indice : cette fois-ci, exceptionnellement dans un QCM, il pourrait y avoir plusieurs bonnes réponses.

Condition de pertinence / rang. A nouveau, on peut et on doit tester la pertinence des instruments. Ici, avec deux instruments à proprement parler Z_1 et Z_2 pour instrumenter un seul régresseur endogène X_1 et sans contrôles, si on écrit la régression linéaire théorique de première étape comme

$$X_1 = \alpha_{01} + \alpha_{02} Z_1 + \alpha_{03} Z_2 + \eta, \text{ où } \mathbb{E}[\eta] = \mathbb{E}[Z_1 \eta] = \mathbb{E}[Z_2 \eta] = 0,$$

la condition de pertinence est équivalente à $(\alpha_{02}, \alpha_{03}) \neq (0, 0)$.

On peut et il faut tester $H_0' : (\alpha_{02}, \alpha_{03}) = (0, 0)$ contre $H_1' : (\alpha_{02}, \alpha_{03}) \neq (0, 0)$ au moyen d'un test de Fisher de significativité jointe dans la régression de première étape.

Condition d'exogénéité. Avec le nouvel instrument Z_2, on a désormais $L = 3 > 2 = K$ (il y a la constante toujours) – deux instruments à proprement parler pour instrumenter un seul régresseur endogène.

On peut alors faire un test de sur-identification de l'hypothèse d'exogénéité des instruments $\mathbb{E}[Z\varepsilon] = 0$ où $Z = (1, Z_1, Z_2)'$ contre l'alternative $\mathbb{E}[Z\varepsilon] \neq 0$ tel que décrit aux slides 32 et 33 du Chapitre 1 d'EM2.