Université Joseph KI-ZERBO

Institut Superieur des Sciences de la Population (ISSP)

Département de Statistique

Exposé d'économétrie

Modèle linéaire à variables instrumentales

Auteurs : **GOUBA Leyla, KABORE Adeline, YAMEOGO Saïdou**

Enseignant : **Dr. Israël SAWADOGO**

Année universitaire : 2024–2025

Table des matières

Introduction

En économétrie, l'objectif principal d'un modèle de régression est d'estimer l'effet causal d'une ou plusieurs variables explicatives sur une variable dépendante. Cependant, cet objectif peut être compromis par la violation de l'hypothèse d'exogénéité, c'est-à-dire une corrélation entre au moins une variable explicative et le terme d'erreur du modèle. Ce phénomène, connu sous le nom d'endogénéité, constitue un obstacle majeur à l'estimation fiable des relations causales. Pour surmonter cette difficulté, les économètres ont recours aux modèles à variables instrumentales (VI).Ces modèles permettent d'identifier et d'estimer les effets causaux même en présence d'endogénéité, à condition de disposer d'instruments valides. Dans ses travaux sur l'impact de l'éducation sur les salaires, l'économiste Joshua Angrist a déclaré : « Un bon instrument est comme une expérience aléatoire que la nature nous offre ». Cette phrase résume l'essence des variables instrumentales (VI), une méthode statistique puissante pour identifier des relations causales dans des données observationnelles, là où les expériences contrôlées sont impossibles.

1 Problématique de l'endogénéité

1.1 Modèle Linéaire Simple : Rappels

Le modèle de régression linéaire simple s'écrit :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \tag{1}$$

Où

$$Y = \beta X + \varepsilon \tag{2}$$

La méthode des Moindres Carrées Ordinaires (MCO) consiste à estimer les paramètres $\beta_0, \beta_1, \beta_2, ...\beta_k$ de telle sorte que $\sum \varepsilon_i^2$ soit Minimale.

$\hat{\beta} = (X'X)^{-1}X'y$

Cette méthode se répose sur 6 hypothèses clés :

— **Linéarité** : La relation entre X et Y est linéaire.

— **Matrice plein rang et absence d'autocorrélation des variables** : $Rang(X) = k$ => $X'X$ est inversible. Cette Propriété garantie l'existence de $\hat{\beta}$

— **Homoscédasticité** : $\text{Var}(\varepsilon|X) = \sigma^2$.

— **Normalité des erreurs** : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

— **Absence d'autocorrélation des résidus** : $Cov(\varepsilon_i, \varepsilon_j X) = 0$ pour $i \neq j$.

— **Exogénéité stricte : absence de corrélation entre X et** ε : $\mathbb{E}(\varepsilon_i|X) = 0$ <=> $\text{Cov}(X, \varepsilon) = 0$

Cette dernière condition est **fondamentale** pour que $\hat{\beta}_{MCO}$ soit **non biaisé**. La violation de cette hypothèse fait perdre

à la Méthode de MCO sa crédibilité car l'estimateur du paramètre est biaisé et non convergent ce qui ne permet de prendre une décision à l'issue de cette régression. **Ce problème est appélé l'endogénéité.**

1.2 Sources d'endogénéité

Comme déjà mentionner plus haut, l'endogénéité survient lorsque . $\text{Cov}(X, \varepsilon) \neq 0$ Ces problèmes peuvent conduire à se tromper dans l'interprétation des paramètres.

Il existe trois types d'endogénéités :

— **Cas 1 : Erreur de mesure sur une variable explicative.** En effet le bruit introduit par l'erreur de mesure réduit la corrélation apparente entre la variable explicative et la variable dépendante. Le coefficient estimé est sous-estimé en valeur absolue (biais d'atténuation). Exemple : Supposons qu'on veux mesurer l'effet du temps d'étude X^* sur la note d'un étudiant Y. Mais le temps d'étude est mal mesuré (auto-déclaré). Si les étudiants exagèrent ou sous-estiment le temps => $X = X$ +v Résultat : l'effet estimé de X sur Y est plus faible que dans la réalité (on sous-estime le paramètre).

— **Cas 2 :Variable omise corrélée avec une variable explicative** Il nous manque une variable explicative importante corrélée positivement à la variable dépendante et cette variable explicative est corrélée fortement et positivement (resp. négativement) à l'une de nos variables explicatives. En effet Si Z (la variable omise) est corrélée avec X, alors X est corrélée avec ε car Z est incluse dans ε : l'existence d'endogénéité. Consequence on surestime (resp. sous-estime) la valeur absolue du paramètre.

Exemple : On estime l'effet de l'éducation X sur le salaire Y, mais on oublie la motivation Z. Z influence à la fois l'éducation (plus motivé = plus d'années d'études) et le salaire (plus motivé = mieux payé). Si on ne contrôle pas Z, alors l'effet de X sera surestimé : on attribue à l'éducation un effet qui vient en réalité de la motivation.

— **Cas 3 : Simultanéité (causalité bilatérale)** la variable explicative dépend par ailleurs positivement de la variable expliquée. Exemple : Étudions l'effet du prix P sur la quantité demandée Q.

Nous savons que le prix est aussi influencé par la demande : si la demande augmente, les producteurs augmentent les prix. Il y a donc un modèle simultané :

$$\mathbf{Q_d} = \alpha - \beta \mathbf{P} + u \tag{3}$$

(demande)

$$\mathbf{P} = \gamma - \delta \mathbf{Q_d} + v \tag{4}$$

(offre)

Utiliser MCO ici donnerait des coefficients biaisés, car le prix est endogène.

Le problème avec l'une de ces situations est que la régression linéaire simple qui pourrait normalement être utilisée dans l'analyse peut produire des estimations incohérentes ou biaisées, c'est là que les variables instrumentales (IV) seraient alors utilisées.

2 Domaines d'application du modèle à variables instrumentales (VI)

1. Économie de l'éducation

Contexte :

— **Objectif** : Mesurer l'effet de l'éducation (années de scolarité) sur les revenus ou les salaires.

— **Problème** : Les individus qui poursuivent leurs études plus longtemps peuvent être naturellement plus motivés, plus intelligents ou issus de familles favorisées. Ces facteurs non observés faussent l'estimation.

Solution par VI :

— **Variables instrumentales possibles :**

  — Distance au collège ou à l'université

  — Réformes éducatives (ex. : allongement obligatoire de la scolarité)

  — Âge minimum de sortie d'école modifié par une loi

— **Pourquoi VI** : Ces instruments influencent l'éducation sans affecter directement le revenu (sauf via l'éducation), ce qui satisfait les conditions de validité.

2. Économie du travail

Contexte :

— **Objectif** : Évaluer l'impact des heures de travail ou du statut d'emploi sur la productivité ou le bien-être.

— **Problème** : Les travailleurs les plus productifs peuvent s'auto-sélectionner vers des emplois avec plus d'heures, rendant la relation biaisée.

Solution par VI :

— **Instruments typiques :**

  — Réduction légale du temps de travail (comme les 35 heures en France)

  — Variations régionales dans la réglementation du travail

— **Pourquoi VI** : Ces politiques affectent les heures de travail mais sont considérées comme exogènes à la productivité individuelle.

3. Santé publique

Contexte :

— **Objectif** : Estimer l'effet d'un traitement médical ou d'une couverture d'assurance sur l'état de santé.

— **Problème** : Les individus qui choisissent de se faire soigner peuvent avoir des comportements ou des niveaux de santé initiaux différents.

Solution par VI :

— **Instruments typiques :**

- Assignation aléatoire dans des expérimentations naturelles (ex. essais cliniques randomisés)
- Changements dans l'éligibilité à l'assurance (ex. réforme Medicare aux États-Unis)

— **Pourquoi VI** : Ces variables modifient l'accès au traitement sans être directement liées à l'état de santé non observé.

4. Économie du développement

Contexte :

— **Objectif** : Analyser l'effet de l'accès à des services comme le microcrédit ou l'irrigation sur le revenu ou la productivité.

— **Problème** : Les zones qui reçoivent un microcrédit peuvent être plus dynamiques ou ciblées selon des critères spécifiques.

Solution par VI :

— **Instruments possibles :**

- Distance au bureau de microcrédit
- Date d'ouverture de l'antenne locale
- Programmes pilotes dans certaines régions

— **Pourquoi VI** : Ces instruments influencent l'accès mais pas directement le revenu, sauf via l'accès.

5. Évaluation des politiques publiques

Contexte :

— **Objectif** : Mesurer l'effet de politiques sociales, d'aides publiques, ou de programmes de formation.

— **Problème** : Les bénéficiaires ne sont pas sélectionnés au hasard — ils peuvent être auto-sélectionnés (plus motivés, en plus grande difficulté, etc.).

Solution par VI :

— **Instruments typiques :**

- Critères d'éligibilité fondés sur le revenu ou l'âge
- Seuils d'attribution (design discontinu)
- Loteries ou quotas d'accès à un programme

— **Pourquoi VI** : Ces dispositifs créent une variation exogène dans le traitement, assimilable à une expérience naturelle.

3 Cadre conceptuel du modèle instrumental (VI)

Soit le modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \tag{5}$$

Encore noté

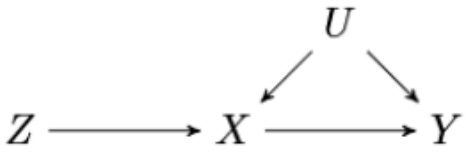
$$Y = \beta X + \varepsilon \tag{6}$$

où ε est d'espérance nulle et de variance condition- nelle égale à σ^2. Ce modèle reprend les notations habituelles mentionnées dans le rappel. Le problème étudié est le suivant : nous soupçonnons une variable explicative d'être endogène. Supposons qu'il s'agisse de la dernière variable du modèle, notée X_k. Dans ce cas, la variable X_k est corrélée avec le terme d'erreur et l'hypothèse $E(X_{ki}, \varepsilon_i) = 0$ n'est plus vérifiée. Ainsi si on applique le MCO sur le modèle (6) qui contient une variable explicative endogène on obtient des paramètres estimés biaisés et inconsistants. On montre que dans ce cas, on peut obtenir des estimateurs consistants grâce à la méthode des variables instrumentales qui est une méthode d'estimation en information limitée c'est à dire que nous nous intéressons à une seule équation. *On appelle variable* ***instrumentale****, une variable exogène (non corrélée avec le terme d'erreur) et corrélée avec la variable explicative qui est soupçonnée d'être endogène ici* X_k.

Un **instrument** est une variable auxiliaire qui permet d'étudier la rélation causale entre deux autres variables. Dans la plupart des cas,l'instrument est une cause antécédente de la variable explicative. Par exemple, pour étudier la relation entre une cause X_k et un effet Y, nous pourrions exploiter l'instrument $Z -> X_k -> Y$. Une bonne variable instrumentale remplit trois conditions : inclusion, exclusion et monotonicité.

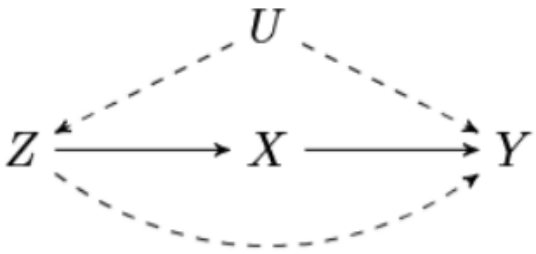
— **Condition 1 : inclusion**

Pour qu'une variable soit un instrument valide, il faut qu'elle soit associée à la variable explicative qui nous intéresse. La condition d'inclusion est satisfaite lorsque l'instrument Z est associé à la variable explicative X. Soit **U** la variable inobservée (incluse dans ε causant $E(X_{ki}, \varepsilon_i) = 0$. Dans le graphe ci-dessous, la variable instrumentale Z remplit la condition d'inclusion, puisqu'elle cause X



— **Condition 2 : exclusion**

Pour qu'une variable instrumentale soit valide, il faut qu'elle soit associée à la variable dépendante seulement à travers la variable explicative. Dans le graphe suivant, la variable instrumentale est associée à la variable dépendante à travers la variable explicative.



Malheureusement, les deux chemins pointillés sont problématiques. D'abord, une variable U cause Z et Y ; cette fourchette laisse circuler l'information statistique entre Z et Y, ce qui viole la condition d'exclusion. Ensuite, l'instrument cause directement la variable dépendante. À moins que l'analyste puisse bloquer les deux chemins pointillés à l'aide de variables de contrôle,Z n'est pas un instrument valide.

En général, la condition d'exclusion est difficile à satisfaire. De plus, contrairement à la condition d'inclusion, la condition d'exclusion est une propriété essentiellement théorique ; les tests empiriques ne peuvent habituellement pas démontrer qu'elle est satisfaite. Les conditions d'inclusion et d'exclusion peuvent être réexprimées de façon plus générale en termes de graphes orientés acycliques. Z est un instrument valide pour estimer l'effet de X sur Y si deux conditions sont satisfaites (Brito et Pearl, 2002) :

— **1. Il existe un chemin ouvert entre Z et X**.

— **2. Tous les chemins ouverts entre Z et Y comprennent une flèche qui pointe vers X** .

— **Condition 3 : monotonicité**

Afin que les résultats de l'analyse par variable instrumentale puissent être interprétés en termes causaux, il faut qu'une troisième condition technique soit remplie : la monotonicité. Cette conditionstipule qu'aucun des participants de l'étude ne doit être anticonfor-miste. Ici, le terme *« anticonformiste »* signifie qu'aucun participantne doit répondre de façon opposée à l'effet moyen de la variable instrumentale sur la variable explicative (Imbens et Angrist, 1994 ; Hernánet Robins, 2020 ; Sovey et Green, 2011).

Par ailleurs il existe une condition fondamentale sur les dimensions de la matrice Z : Soit **Z** la matrice des variables instrumentales de dimension **(N,p)** avec **p k** c'est-à-dire que le nombre d'instruments doit être supérieur ou égal au nombre de variables explicatives endogènes du modèle.

— **Si p = k le modèle est juste identifié**

— **si $p > k$ le modèle est suridentifié.**

Attention ! si $p < k$ le modèle est sous identifié et on ne peut pas retrouver les paramètres structurels.

4 Méthode d'estimation

Soit le modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \tag{7}$$

Ou

$$Y = \beta X + \varepsilon \tag{8}$$

On suppose que X_k est **endogène** (corrélée avec l'erreur ε).

Posons $Z = (1, X_1, X_2, \ldots, X_{k-1}, Z_1, \ldots, Z_m)$ avec $p = k + m$ où p est le nombre de colonnes de Z et $Z_1, \ldots, Z_m$ les variables instrumentales.

Remarque : les notations peuvent prêter à confusion ; ici la matrice Z contient toutes les variables exogènes, c'est-à-dire d'abord toutes les variables exogènes du modèle ou encore $1, X_1, X_2, \ldots, X_{k-1}$ ainsi que toutes les variables instrumentales $(Z_1, \ldots, Z_m)$.

Remarquons de plus que les variables instrumentales Z_i ne figurent pas dans l'équation structurelle ; ces restrictions d'exclusion permettent d'identifier le modèle .

On montre dans ce cas que, pour obtenir l'estimateur des Variables Instrumentales, il faut procéder selon les deux étapes suivantes :

Étape 1 : Régression auxiliaire (prédiction de X_k) : On pose le modèle suivant

$$X_k = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{k1} X_{k1} + \delta_1 Z_1 + + \delta_m Z_m + v_i \tag{9}$$

On regresse X_k sur $(1, X_1, X_2, \ldots, X_{k-1}, Z_1, \ldots, Z_m)$ par la méthode de MCO et on calcule $\hat{X}_k$ grâce à cette régression. Notons $\hat{X}$ la matrice qui contient les k colonnes suivantes : $1, X_1, X_2, \ldots, X_{k-1}, \hat{X}_k$ où X_k a été remplacée par $\hat{X}_k$. Remarquons que la régression de cette première étape n'a pas d'interprétation économique. À nouveau elle pourrait être nommée **"régression auxiliaire"** car elle aide à la correction du biais. On peut réecrire le modèle de départ en remplaçant X_k avec $\hat{X}_k$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k \hat{X}_k + \varepsilon \tag{10}$$

Ou

$$Y = \beta \hat{X} + \varepsilon \tag{11}$$

Étape 2 : on regresse Y sur $\hat{X}$, c'est-à-dire sur les variables $(1, X_1, X_2, \ldots, X_{k-1}, \hat{X}_k)$. Les paramètres estimés à la fin de cette seconde étape sont les estimateurs par la méthode des variables instrumentales. Cette estimation en deux étapes correspond aussi à la méthode des **Doubles Moindres Carrés, DMC (ou 2SLS)**. On obtient dans cette seconde étape :

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = \hat{\beta}_{IV} \tag{12}$$

Rappel : X est de dimension (N, k).

5 Test de spécification

La force d'un instrument réside essentiellement dans sa corrélation avec la variable explicative endogène ; un instrument puissant est celui qui est fortement corrélé à la variable endogène mais non corrélé au terme d'erreur dans l'équation structurelle. Ceci est crucial car un instrument faible peut conduire à des estimations de paramètres biaisées et incohérentes, minant ainsi la crédibilité de l'estimation IV.

D'un point de vue économétrique, la pertinence et la validité d'un instrument sont primordiales. La pertinence fait référence à la capacité de l'instrument à expliquer la variation du régresseur endogène, tandis que la validité concerne l'absence de corrélation directe entre l'instrument et le terme d'erreur. Ces deux conditions sont souvent évaluées au moyen de divers tests et critères statistiques, tels que la **statistique F** dans la régression de première étape et le **test de Sargan-Hansen** pour la suridentification des restrictions.

Test d'endogénéité : Éléments théoriques

Source : Wooldridge 2006, p. 532.

Le test d'endogénéité permet de vérifier si une ou plusieurs **variables explicatives** dans un modèle de régression sont **corrélées avec l'erreur du modèle**. Si c'est le cas, ces variables sont dites **endogènes**, ce qui rend les estimateurs MCO (Moindres Carrés Ordinaires) **biaisés et inconsistants**. On a alors besoin d'utiliser des **variables instrumentales** et la méthode des doubles moindres carrés (**2SLS**) . L'idée du test est que la différence entre les estimateurs des MCO et des 2SLS doit être faible si la variable explicative est exogène. Si cette différence est « grande », on conclut que la variable explicative suspectée est endogène. Pour savoir si cette différence est faible, on peut utiliser une régression avec le raisonnement suivant : supposons que nous ayons une seule variable explicative suspectée d'être endogène, notée Y_2 dans la suite. Le modèle structurel est le suivant :

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 X_1 + \beta_3 X_2 + u_1. \tag{13}$$

On dispose de deux variables exogènes supplémentaires, les deux instruments, notés Z_1 et Z_2.

La forme réduite s'écrit :

$$Y_2 = \pi_0 + \pi_1 X_1 + \pi_2 X_2 + \pi_3 Z_1 + \pi_4 Z_2 + u_2. \tag{14}$$

Si Y_2 n'est pas corrélée avec u_1, nous devons utiliser les MCO. Dans quel cas la corrélation entre Y_2 et u_1 est nulle ? On a :

$$\mathbf{corr}(Y_2, u_1) = \mathbf{corr}(\pi_0 + \pi_1 X_1 + \pi_2 X_2 + \pi_3 Z_1 + \pi_4 Z_2 + u_2, u_1) = \mathbf{corr}(u_2, u_1).$$

Donc la variable Y_2 n'est pas corrélée avec le terme d'erreur u_1 si ce terme d'erreur n'est pas corrélé avec u_2. Comment tester la nullité de cette corrélation ? Écrivons $u_1 = \delta u_2 + \varepsilon$ où ε vérifie toutes les hypothèses des MCO.

Si u_1 et u_2 ne sont pas corrélés alors $\delta = 0$. Il suffit donc de remplacer u_1 par l'expression ci-dessus dans l'équation structurelle pour obtenir :

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 X_1 + \beta_3 X_2 + \delta u_2 + \varepsilon.$$

On remplace u_2 par le résidu de la forme réduite et on teste $\delta = 0$.

*** On peut résumer ce test en deux étapes :**

— **Étape 1 :** On estime la forme réduite par MCO en régressant Y_2 sur Z, c'est-à-dire

$$Y_2 = \pi_0 + \pi_1 X_1 + \pi_2 X_2 + \pi_3 Z_1 + \pi_4 Z_2 + u_2$$

et on récupère les résidus $\hat{u}_2$.

— **Étape 2 :** On inclut $\hat{u}_2$ dans l'équation structurelle :

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 X_1 + \beta_3 X_2 + \delta \hat{u}_2 + \varepsilon$$

et on teste $H_0 : \delta = 0$.

Si H_0 est rejetée, alors Y_2 est endogène et il faut utiliser les estimateurs 2SLS.

Test d'endogénéité de Hausman (Durbin-Wu-Hausman)

Le **test de Hausman** permet de vérifier si certaines variables explicatives sont endogènes, c'est-à-dire corrélées avec le terme d'erreur. En cas d'endogénéité, l'estimateur des moindres carrés ordinaires (MCO) est biaisé et inconsistent, ce qui justifie l'usage de la méthode des variables instrumentales (2SLS).

Hypothèses

— H_0 : Pas d'endogénéité. Les estimateurs MCO et 2SLS sont proches. L'estimateur MCO est convergent.

— H_1 : Présence d'endogénéité. Les estimateurs MCO et 2SLS sont significativement différents. L'estimateur MCO est biaisé et inconsistent.

Étapes du test

1. **Estimation par MCO** : On estime le modèle linéaire avec les variables potentiellement endogènes.

$$Y = X\beta + u$$

$$\hat{\beta}_{\text{MCO}} = (X'X)^{-1}X'Y$$

2. **Estimation par variables instrumentales (2SLS)** : On estime les paramètres suite à la regréssion du modèle instrumental expliqué plus haut.

$$\hat{\beta}_{\text{2SLS}} = (X^{*'}X^{*})^{-1}X^{*'}Y$$

où X^* est la projection de X sur l'espace engendré par les instruments Z.

3. **Calcul de la statistique de Hausman** :

$$H = (\hat{\beta}_{2SLS} - \hat{\beta}_{MCO})' \left[\text{Var}(\hat{\beta}_{2SLS}) - \text{Var}(\hat{\beta}_{MCO})\right]^{-1} (\hat{\beta}_{2SLS} - \hat{\beta}_{MCO})$$

4. **Distribution sous** H_0 :

$$H \sim \chi^2_k$$

où k est le nombre de variables suspectées d'endogénéité.

Décision

— Si la statistique H est **faible** (valeur-p > 5%) : On ne rejette pas H_0. Donc pas d'endogénéité et le modèle de MCO est acceptable.

— Si la statistique H est **élevée** (valeur-p < 5%) : On rejette H_0. Présence d'endogénéité. Préférer l'estimateur 2SLS.

Test de sur-identification (Test de Sargan-Hansen)

Le test de Sargan ou test de Sargan-Hansen est un test statistique permettant de tester une hypothèse de suridentification dans un modèle statistique. Il est aussi connu sous le nom de test de Hansen ou test J. En pratique, on est souvent amené à effectuer des estimations d'une même équation en étendant ou en restreignant la liste des variables instrumentales. En effet, on pouvait avoir intérêt à accroître le nombre de variables instrumentales dans la mesure où cela conduit à des estimateurs plus précis. Cependant, accroître indûment l'ensemble des variables instrumentales peut conduire à faire apparaître des biais dans l'estimation. Lorsqu'on dispose de plus de variables exogènes que nécessaires, on dit que le modèle est suridentifié. Le test de suridentification de Sargan est un test très important et très couramment utilisé permettant de contrôler qu'il n'y a pas d'incohérence dans le choix des variables instrumentales. Ce test, appelé test de suridentification, ou test de Sargan constitue un guide incontournable dans le choix des variables instrumentales.

Il permet de tester la validité d'un instrument et de vérifier son exogénéité. Ce test est valable sous deux conditions :

— **Au moins un des instruments est valable, c'est-à-dire qu'il est exogène.**

— **Il faut plus d'instruments que de variables instrumentées.**

Le test de Sargan est construit sur les hypothèses suivantes :

— **Hypothèse nulle (H) : Validité des instruments** *Tous les instruments sont* ***exogènes****, c'est-à-dire qu'ils ne sont* ***pas corrélés*** *avec le terme d'erreur du modèle.*

— **Hypothèse alternative (H) : Au moins un instrument est endogène** *Il existe au moins un instrument qui est* ***corrélé*** *au terme d'erreur*

Etapes de réalisation du test

1. Estimer le modèle par la méthode des doubles moindres carrés (2SLS).
2. Récupérer les résidus estimés $\hat{u}$.
3. Régréssion auxiliaire : régresser les résidus sur l'ensemble des instruments.
4. Calcul de la statistique :

$$S = N \cdot R^2$$

où :

— N est la taille de l'échantillon,

— R^2 est le coefficient de détermination de la régression auxiliaire.

Sous l'hypothèse nulle H_0 (instruments valides), la statistique S suit une loi du χ^2 avec :

$$\text{ddl} = \text{nombre d'instruments} - \text{nombre de variables endogènes}$$

Décision

— Si la statistique H est **faible** : on ne rejette pas H_0 . Les instruments sont dans ce cas valides (pas de corrélation significative avec les résidus).

— Si la statistique H est **élevée** : on rejette H_0 , il y a au moins un instrument corrélé au terme d'erreur (donc invalide).

Test de Sous-Identification dans les Modèles à Variables Instrumentales

Le test de sous-identification vérifie si les instruments sont suffisamment corrélés avec la variable endogène (condition de pertinence). Il détecte les situations où :

— Les instruments sont faibles (faible corrélation avec X)

— Le modèle est mal spécifié (nombre d'instruments insuffisant)

Statistiques Clés :

— Kleibergen-Paap rk LM (test de rang)

— Cragg-Donald F-statistique

— Anderson canonical correlations LM

Hypothèses :

— H_0 : Modèle sous-identifié (instruments non pertinents)

— H_1 : Instruments pertinents

Implémentation Pratique

Étapes :

1. Faire une régression MCO et DMCO en calculant à chaque fois le R^2
2. Calculer la statistique de test :

F-statistique (Cragg-Donald) :

$$F = \frac{(R^2_{\text{non contraint}} - R^2_{\text{contraint}})/k}{(1 - R^2_{\text{non contraint}})/(n-m)}$$

où k = nombre d'instruments, m = nombre de variables exogènes, n = taille de l'échantillon

Modèle non restreint : Régression de la variable endogène **avec** les instruments.

Modèle restreint : Régression de la variable endogène **sans** les instruments (Modèle de depart).

$$R^2_{\textbf{non contraint}} = 1 - \frac{\sum(y_i - \hat{y}_i^{IV})^2}{\sum(y_i - \bar{y})^2}$$

$$R^2_{contraint} = 1 - \frac{\text{SC}_{\text{résiduelle}}}{\text{SC}_{\text{totale}}} = \frac{\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n(y_i - \bar{y})^2}$$

Statistique LM (Kleibergen-Paap) :

$$LM = n \times R^2_{\text{auxiliaire}} \sim \chi^2(k)$$

TABLE 1 – Règles de décision pour les tests de sous-identification

Test	**Statistique**	**Valeur Critique**	**Conclusion**
Cragg-Donald F	$F > 10$	≥ 10 : Instruments forts	Rejet de H_0 (pertinence confirmée)
Kleibergen-Paap LM	p-value < 0.05	$\chi^2(k)$	Rejet de H_0
Anderson LM	p-value < 0.05	–	Rejet de H_0

Test de pertinence des instruments

La **pertinence** (ou **force**) des instruments est une condition **essentielle** pour la validité des variables instrumentales (VI). Elle exige que les instruments soient **fortement corrélés** avec la variable endogène. Si les instruments sont faibles :

— L'estimateur VI devient **biaisé** et **imprécis**

— Les tests d'hypothèse perdent leur fiabilité

— Le biais peut être pire qu'avec les MCO (problème dit "d'instruments faibles")

1. test de significativité partielle

Il s'agit de tester si le coefficient d'un instrument est **significativement non nul** dans la régression auxiliaire.

Hypothèse : On procède tout d'abord par une régression auxiliaire. Puis on calcul la statistique du test $t_j = \frac{\hat{\pi}_j}{SE(\hat{\pi}_j)}$

$\hat{\pi}_j$ = Coefficient estimé de l'instrument Z_j

SE($\hat{\pi}_j$) = Erreur standard du coefficient

2. test de significativité globale

On tester la significativité **conjointe** des instruments. La statistique est donnée par

$$F = \frac{(SSR_{\text{restreinte}} - SSR_{\text{non restreinte}})/q}{SSR_{\text{non restreinte}}/(n-k-1)}$$

où

SSR = Somme des carrés des résidus

q = nombre d'instruments

k = nombre total de régresseurs (variables explicatives + variables instrumentales)

— **Modèle non restreint** : Régression de la variable endogène **avec** les instruments.

— **Modèle restreint** : Régression de la variable endogène **sans** les instruments.

Règle empirique (Stock-Yogo, 2005) :

- $F < 10$: Instruments faibles (problème sérieux)
- $F > 10$: Instruments acceptables
- $F > 20$: Instruments forts

6 Interprétation des paramètres dans un modèle à variables instrumentales

6.1 Les paramètres du modèle

Considérons le modèle (6) mentionné plus haut. Après la régression à **Double MCO** on obtient

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = \hat{\beta}_{IV} \tag{15}$$

L'interprétation des $\hat{\beta}_i$ n'est valable que si les instruments Z_i sont de bonne qualité. C'est - à - dire :

— **Pertinence** : $Cov(Z_i, X_i) \neq 0$;

— **Exogénéité** : $Cov(Z_i, u_i) = 0$.

Nous supposons que les instruments utilisés sont valides. L'estimateur IV de β_i mesure l'effet **causal** de X_i sur Y, en corrigeant le biais d'endogénéité. Il représente le **LATE** (Local Average Treatment Effect), c'est-à-dire l'effet moyen pour les individus affectés par la variation de l'instrument.

Remarque : L'interprétation des $\hat{\beta}_{IV}$ ne porte **pas nécessairement** sur la population entière mais uniquement sur les variables endogènes de depart.

Exemple : Supposons que Y_i représente le revenu, X_i les années d'éducation, et Z_i la distance à l'école durant l'enfance. Si $\hat{\beta}_1 = 0.09$, on peut interpréter cela comme : une année supplémentaire d'éducation induite par la proximité d'une école augmente le revenu moyen de 9%.

6.2 Le R^2

Le coefficient de détermination R^2 dans un modèle à variables instrumentales (2SLS) n'a pas d'interprétation directe en termes de proportion de variance expliquée comme en MCO, en raison de la correction de l'endogénéité. Toutefois :

— **Le R^2 de la première étape** (c'est - à - dire le R^2 de la regression auxiliaire) donné par

$$R^2 = \frac{\sum(\hat{X}_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}$$

est utile pour évaluer la **pertinence des instruments**. Un R^2 élevé indique que les instruments expliquent bien la variable endogène.

— **Le R^2 de la seconde étape** donné par

$$R^2_{\text{2SLS}} = 1 - \frac{\sum(y_i - \hat{y}_i^{IV})^2}{\sum(y_i - \bar{y})^2}$$

ne doit pas être utilisé pour juger de la qualité de l'ajustement du modèle. Il peut même être **négatif** , car l'estimateur IV n'est pas obtenu par minimisation de la somme des carrés des résidus comme en MCO.

TABLE 2 – Comparaison du coefficient de détermination R^2 entre les modèles MCO et IV

Caractéristique	**Modèle MCO (OLS)**	**Modèle à variables instrumentales (2SLS)**
Définition de R^2	$R^2 = 1 - \frac{SC_{\text{résiduelle}}}{SC_{\text{totale}}}$	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i^{IV})^2}{\sum(y_i - \bar{y})^2}$ (pseudo-R^2)
Interprétation	Proportion de la variance de y expliquée par X	Pas de véritable interprétation comme qualité d'ajustement
Utilité	Mesure directe de la qualité du modèle	Faible utilité ; attention à l'interprétation
Pertinence pour juger les instruments	Pas concerné	Le R^2 de la **première étape** du 2SLS mesure la pertinence des instruments
Remarque	$R^2 \in [0, 1]$, plus il est proche de 1, meilleur est l'ajustement	Peut être faible même avec des instruments valides ; ne reflète pas la consistance des estimateurs

6.3 Limites de l'interprétation

— L'interprétation est **locale** c'est - à- dire valable pour les variables endogènes ;

— L'estimation est sensible à la qualité des instruments :

— Trop d'instruments peut introduire des biais (voir test de Sargan).

Conclusion

Le modèle linéaire avec variables instrumentales constitue une approche puissante pour surmonter les problèmes d'endogénéité souvent rencontrés dans les analyses économétriques appliquées. Contrairement au modèle des moindres carrés ordinaires (MCO), il permet d'identifier l'effet causal d'une variable explicative même lorsqu'elle est corrélée à l'erreur du modèle, ce qui garantirait des estimations biaisées dans un cadre standard.

À travers une méthode d'estimation rigoureuse (comme la méthode des doubles moindres carrés) et des tests de spécification (test de sur-identification, test de sous-identification), le modèle VI assure la validité et la pertinence des instruments utilisés. Il permet ainsi d'isoler la composante exogène de la variable endogène, ce qui renforce la crédibilité des résultats obtenus.

Ce cadre s'avère particulièrement pertinent dans de nombreux domaines tels que l'économie de l'éducation, la santé publique, le développement ou l'évaluation de politiques publiques. Il est cependant crucial de bien sélectionner les instruments, car la validité de l'approche repose entièrement sur leur pertinence (corrélation avec la variable endogène) et leur exogénéité (absence de lien avec l'erreur du modèle structurel).

Références

[1] Variables instrumentales informations instrumentales regression probit : https ://fastercapital.com/fr/contenu/Variables-instrumentales—informations-instrumentales—regression-probit-et-variables-instrumentales.html

[2] FOAD :https ://cmaurel.wordpress.com/wp-content/uploads/2013/05/master-foad-chap-3-var-instr.pdf

[3] Chapitre 7 les variables instrumentales :https ://www.academia.edu/35542566/Chapitre$_{7}Les_{V}ariables_{I}nstrumentales$

[4] Les variables instrumentales : https ://www.scribd.com/document/647549414/Les-Variables-Instrumentales

[5] Program Evaluation Instrumentals Variables :https ://evalsp24.classes.andrewheiss.com/example/iv.html ?utm$_{s}ource = chatgpt.com$