

Econométrie 2, 2018-2019

Consignes et sujets

Jérémy L'Hour
assistant-econometrie@ensae.fr

March 15, 2019

1. Consignes

Mode de fonctionnement, attribution, notation. Le projet est à effectuer et rendre par binôme d'étudiants d'un même groupe de TD. Lorsque les groupes sont de taille impaire, et une fois le plus possible de binômes formés, il est possible de former un groupe de trois étudiants.

Il vous est demandé d'envoyer le numéro de vos trois sujets préférés à votre chargé de TD pour le lundi 18 mars à 18h. L'attribution des sujets répond aux contraintes suivantes: (i) chaque sujet doit être pris une fois dans un groupe de TD, (ii) un sujet peut être pris deux fois au plus dans un groupe de TD. C'est à chaque chargé de TD de déterminer la répartition des sujets, en essayant de répondre aux préférences des élèves. C'est le concepteur du sujet qui va corriger tous les projets réalisés à partir de son sujet. La triche et le plagiat seront sévèrement sanctionnés.

Date de rendu. Vendredi 17 mai, 18h: **réponses rédigées ET code.**

Notation. Le sujet est noté sur 20 points. Le langage de programmation est au choix parmi R, STATA, SAS ou Python.

Liste des sujets.

Num.	Intitulé	Rendre à
1	Position sur le marché du travail et inégalités liées à l'origine	lucas.girard @ensae.fr
2	Influence d'une chaîne de TV conservatrice sur le vote	jeremy.l.hour @ensae.fr
3	Impact de l'arrestation sur la probabilité de récidive	jeremy.l.hour @ensae.fr
4	La beauté et la couleur de peau en politique	o.couperier @gmail.com
5	Impact du salaire minimum dans la restauration rapide	yannick.guyonvarch @ensae.fr
6	Estimation des rendements privés et sociaux de l'enseignement supérieur	elio.nimier-david @ensae.fr
7	Création d'un indice de prix pour l'immobilier commercial	roxane.morel @developpement-durable.gouv.fr
8	Fécondité et participation des femmes au marché du travail en Afrique sub-saharienne	raphael.sh.lee @gmail.com

2. Pièges à éviter

2.1. Utilisation des méthodes

- Tout résultat d'estimation doit faire figurer un écart-type dont on précisera le modèle de calcul (niveau de cluster, par exemple).
- La matrice de corrélation des régresseurs ne permet PAS d'étudier l'hypothèse de non-corrélation entre régresseurs et résidus. La conclusion "il y a peu de corrélation entre régresseurs donc on peut penser qu'il n'y a pas de gros problème d'endogénéité" est fausse. L'hypothèse d'exogénéité est difficilement testable et repose sur une justification soigneuse avant toute estimation.
- En revanche, si les variables sont très corrélées, les estimateurs peuvent être très peu précis.
- Le fait que les X soient corrélés avec le Y ne permet pas de conclure à l'existence d'endogénéité (corrélation entre X et ε).
- Quand vous faites un Probit ou un Logit, les coefficients ne s'interprètent pas directement. Il faut passer par les effets marginaux.
- Attention à bien calculer les effets marginaux quand une variable et sa transformation sont dans le modèle. Attention par exemple pour les calculs d'effets marginaux avec age et age au carré où age au carré apparaît car il a été calculé à la main puis intégré dans le modèle.
- Les coefficients dans les modèles de type Logit ou Probit ordonnés avec seuils connus ont la même interprétation que dans une régression linéaire classique (c'est simplement une régression par intervalles due au fait que l'information n'est que partielle).
- Faites toujours figurer les écarts-types dans vos tableaux de résultats, les étoiles ne sont pas suffisantes.

- L’utilisation d’un modèle binomial de type Logit ou Probit ne se justifie PAS pas le fait que l’un des X soit binaire mais par le fait que Y soit binaire.
- Y compris entre 0 et 1 ne justifie par l’utilisation du modèle binomial : il faut que Y vaille SOIT 0, SOIT 1.
- Dans les modèles polytomiques non ordonnés, il y a toujours une référence. Par exemple dans un modèle de choix de transport, si c’est la voiture personnelle qui aurait été normalisée, alors les coefficients associés aux autres modalités s’interpréteront par rapport à l’alternative “voiture personnelle”.
- L’hypothèse IANP (Indépendance aux Alternatives Non Pertinentes) est rarement mentionnée et encore moins discutée dans les modèles de type “multinomial Logit”. Il est important de le faire.
- Adopter le réflexe d’écrire le modèle et ses hypothèses. La vraisemblance aussi, si c’est la méthode choisie.
- Attention au traitement des valeurs manquantes dans vos données. Elles sont parfois codées à 0 par inattention ce qui peut poser problème dans la suite de vos estimations. Une justification de la méthode de traitement est nécessaire (e.g. suppression, remplacement par la valeur la plus fréquentes, remplacement en utilisant un algorithme de type plus-proche-voisin). Eventuellement: est-ce que la prise en compte de la sélection est nécessaire dans le problème ?
- N’incluez pas des variables endogènes quand elles ne vous intéressent pas.

2.2. Présentation, problèmes de forme

- Utilisez le nom français de la quantité mesurée dans vos phrases: votre variable sera le nombre d’année d’études et non pas *nb_an_educ*.
- Faire un copier-coller de sortie STATA n’est pas acceptable : il a des packages qui permettent de sortir un beau tableau LaTeX sans aucun effort.

- Les graphiques doivent porter un message : réfléchissez à la meilleure façon de représenter ce message et rendez-le facilement compréhensible pour le lecteur. Les statistiques descriptives doivent déjà vous donner une idée des variables pertinentes à inclure et sont l’occasion de démontrer vos compétences en visualisation des données.
- Pensez au sens des phrases que vous écrivez. Personne ne passe de la catégorie “homme” à “femme” pour le sexe; on n’est pas “célibataire” puis tout à coup “divorcé”.
- Au même titre que vous devez faire attention à l’échelle de mesure de vos variables, vous devez aussi préciser les unités de mesure. On n’écrit pas “ un salaire de 3000 ” sans préciser que ce sont des euros.

Projet Econométrie 2

Position sur le marché du travail et effets de l'origine (nationalité des parents)

Chargé de TD : Lucas Girard – lucas.girard@ensae.fr

Présentation Ce projet cherche à mesurer l'effet de différentes variables individuelles sur la position sur le marché du travail, étudiée ici par une marge extensive (actif occupé, chômeur, inactif) et une marge intensive (salaire horaire). La principale variable explicative d'intérêt est une variable catégorielle "origine" construite à partir des nationalités des parents, à leurs naissances.¹ Cette variable pourra également être considérée en interaction avec d'autres variables, notamment la variable "immi" qui est l'indicatrice d'être immigré (être né étranger à l'étranger).

L'objectif principal du projet consiste à mesurer l'effet de la variable "origine" afin de réfléchir à l'existence d'éventuelles inégalités dues à l'origine des individus sur le marché du travail.

Consignes générales Le projet n'est pas fondé sur un article précis et présente un caractère exploratoire. Il laisse une certaine liberté dans le choix des modèles estimés mais nécessite en contrepartie d'être extrêmement précis sur : (i) les observations utilisées, (ii) les variables incluses dans l'analyse, (iii) la méthode d'estimation et d'inférence (calcul des écarts-types). Une attention particulière sera portée aux choix des modèles retenus, à la discussion de leurs hypothèses, aux paramètres d'intérêt considérés et à l'interprétation des estimations obtenues.

Le projet comporte trois objectifs pédagogiques :

1. appliquer des méthodes vues en cours (variables instrumentales et données de panel) ;
2. illustrer la difficulté à répondre de manière justifiée et suffisamment complète à une question empirique – existe-t-il des inégalités sur le marché du travail liées à l'origine des individus – en combinant données et modèles économétriques ;
3. découvrir deux méthodes d'estimation classiques en données panel lorsque la variable dépendante est binaire ("random effect probit" et "fixed effect logit").²

Données La base de données comporte $n = 70,944$ individus suivis pendant $T = 6$ trimestres, entre 2014 et 2016 (panel cylindré). Pour chaque observation (individu \times trimestre), on dispose de variables :

- démographiques et socio-économiques individuelles : âge, sexe, diplôme, CSP des parents, "origine", "immi", une indicatrice d'être descendant d'immigré ("desc"), état de santé déclaré, statut matrimonial, etc.
- individuelles liées à la position sur le marché du travail : statut (actif occupé, chômeur, inactif), salaire, nombre d'heures travaillées, type d'horaires, type de contrat de travail, etc.
- relatives à la zone résidentielle dans laquelle habite l'individu : tranche d'unité urbaine, commune urbaine ou rurale, proportions de populations de différentes origines dans cette zone, etc.

1. Il s'agit d'un *proxy* fréquemment utilisé pour mesurer le groupe ethnique ou l'origine d'un individu. La variable "origine" prend ici neuf modalités : France, Europe du Nord, Europe du Sud, Europe de l'Est, Maghreb, reste de l'Afrique, Proche-Orient, Asie du Sud-Est (Cambodge, Laos, Vietnam), reste du monde.

2. Des références sont fournies ; cette partie théorique consiste essentiellement à découvrir et comprendre ces deux méthodes et à les appliquer.

1 Marge intensive (salaire net horaire)

Dans un premier temps, on étudie le salaire des actifs occupés et on s'intéresse à l'existence de potentielles inégalités salariales dues à l'origine.³ Le salaire n'est disponible qu'au premier et au dernier trimestre : données de panel avec $T = 2$ pour cette variable.

Le concept d'inégalité salariale due à l'origine est similaire à celui d'inégalité salariale homme-femme. Il existe un lien étroit avec l'approche économétrique puisqu'on parlera ici d'inégalité salariale due à l'origine lorsque que *toutes choses égales par ailleurs* les salaires des individus diffèrent selon leur origine. L'idée étant que dès lors que ces différences de salaire ne sont pas justifiées (puisque'on se place "toutes choses égales par ailleurs"), elles peuvent être interprétées comme des inégalités.⁴ La question cruciale concerne donc les variables explicatives incluses dans le "toutes choses égales par ailleurs". Au cours du projet, vous aurez justement une certaine liberté dans le choix des variables utilisées.

Afin de vous guider dans votre réponse à la problématique, vous traiterez les questions suivantes.

Question 1 On néglige pour l'instant la dimension panel en utilisant uniquement le dernier trimestre observé pour les individus actifs occupés. On considère un modèle linéaire avec comme variable dépendante le salaire net horaire (variable "salhoraire") ou le logarithme du salaire net horaire (variable "logsalhoraire").

- Estimer ces deux modèles pour un choix de variables explicatives et expliquer vos choix. Différentes spécifications peuvent être essayées et présentées. On discutera des problèmes potentiels d'endogénéité et des distinctions entre corrélation et causalité.
- Interpréter quantitativement et qualitativement les coefficients estimés pour les variables explicatives d'intérêt retenues, lorsque la variable dépendante est le salaire net horaire et lorsque la variable dépendante est le logarithme du salaire net horaire.⁵
- Par quels canaux passe l'effet de la variable "origine" sur le salaire ? On pourra regarder l'évolution des coefficients estimés de la variable "origine" selon les variables explicatives incluses dans le modèle : éducation, types de contrat de travail, secteurs d'activités, etc.
- Quel peut-être l'intérêt de cette analyse (question 1.c) en termes de politiques publiques ?

Question 2 Un déterminant important du salaire est le niveau de diplôme et on souhaite dans cette question l'inclure comme variable de contrôle.

- Quel problème pourrait survenir en incluant cette variable explicative dans les modèles estimés à la question 1 ? Quel serait l'effet en particulier sur l'estimation du coefficient de la variable explicative "origine" ?
- Parmi les variables disponibles dans la base, proposer des variables permettant d'instrumenter le niveau de diplôme et discuter de leur validité. On sera attentif au nombre d'instruments nécessaires.
- Estimer le modèle correspondant où le niveau de diplôme est instrumenté par les variables déterminées à la question précédente. La condition de rang est-elle vérifiée ?
- Interpréter les résultats obtenus au regard de la problématique. Vous donnerez en particulier une interprétation quantitative du coefficient estimé pour la variable explicative "origine".

3. On néglige donc l'aspect censuré de la variable salaire en se restreignant aux individus pour lesquels le salaire est observé (actifs occupés).

4. Ces explications visent à préciser la problématique de ce projet. Elles n'épuisent évidemment pas la question des inégalités qui fait intervenir de multiples dimensions : égalité contre équité, situation à un instant donné versus analyse dynamique tout au long de la vie d'un individu, etc.

5. Il est au moins attendu de considérer l'effet de la variable explicative "origine" mais on pourra également considérer d'autres variables construites à partir des variables "origine", "immi" et "desc" (par exemple : interactions, regroupement de plusieurs modalités de la variable "origine").

Question 3 Sous certaines hypothèses, les panels peuvent également résoudre un problème d'endogénéité.⁶ Contrairement aux deux questions précédentes, on utilise ici la structure de panel des données pour les variables relatives aux salaires : on considère les observations au premier et au dernier trimestre pour les individus actifs occupés.

- a) Estimer un modèle où l'on suppose l'exogénéité des résidus et des effets fixes individuels.⁷
- b) Est-il possible d'inclure la variable "origine" dans le modèle précédent (question 3.a) ? Ce modèle permet-il de résoudre de potentiels problèmes d'endogénéité ?
- c) On autorise désormais les effets individuels à être corrélés avec les régresseurs. Est-il possible dans ce cadre d'étudier l'effet de la variable "origine" ? Le cas échéant, estimer le modèle correspondant et interpréter les résultats obtenus.

2 Marge extensive (actif occupé *vs.* chômeur)

Les inégalités sur le marché du travail dues à l'origine pourraient également survenir sur une marge extensive : avoir ou non un emploi. On s'intéresse dans cette partie aux individus actifs (occupés ou chômeurs) avec comme variable dépendante le fait d'être ou non au chômage. Vous avez à nouveau une certaine liberté quant aux choix des variables explicatives, qu'il faudra justifier, sachant qu'on s'intéresse en particulier, lorsque cela est possible, à l'effet de la variable "origine". Le statut sur le marché du travail (actif occupé, actif chômeur, inactif) est disponible chaque trimestre (données de panel avec $T = 6$).

Question 4 La variable dépendante est binaire. On néglige pour l'instant cette limitation et on la traite comme une variable continue. On utilise donc dans cette question un modèle de probabilité linéaire et on prend en compte la structure de panel des données.

- a) Estimer un modèle de panel où l'on suppose l'exogénéité des résidus et des effets fixes individuels.
- b) Discuter la crédibilité des hypothèses d'exogénéité stricte et d'exogénéité faible et estimer un modèle de panel où l'on autorise la corrélation entre les effets individuels et les régresseurs. Quel problème rencontre-t-on dans ce cadre pour étudier l'effet de la variable "origine" ?

On prend désormais en compte le fait que la variable dépendante est binaire.

Question 5 On se place dans cette question dans le cas "exogénéité des résidus mais autocorrélation" de votre cours sur les panels.

- a) On s'intéresse aux conséquences d'avoir une variable dépendante binaire dans ce cadre : les hypothèses sur les résidus dans les modèles probit ou logit sont-elles compatibles avec l'existence d'un effet individuel et d'une autocorrélation des erreurs ?
- b) Sur les données de panel, estimer un modèle probit (ou logit) avec pour variable dépendante le fait d'être au chômage. Interpréter les estimations obtenues, notamment pour la variable explicative "origine". Les estimations obtenues sont-elles crédibles dans ce cas ?
- c) Estimer le même modèle probit (ou logit) mais avec des données "cross-section" i.e. sans dimension panel et comparer les estimations obtenues dans les deux cas.⁸

6. Pour toutes les questions sur les panels, on gardera bien en tête la structure du cours et la distinction entre : le cas où l'on suppose l'exogénéité des termes individuels inobservés, notés α_i dans le cours (cadre souvent appelé "random effects") ; le cas où l'on autorise les termes individuels à être corrélés avec les régresseurs (cadre "fixed effects").

7. Vous pourrez reprendre les mêmes spécifications que celles utilisées et discutées dans les questions 1 et 2.

8. Pour se ramener à des données "cross-section", on pourra prendre un trimestre de référence arbitraire ou chercher à synthétiser les six trimestres, avec la modalité la plus fréquente par exemple.

Question 6 Dans la section "exogénéité stricte" du cours sur les panels, vous avez vu deux transformations (*first-difference* et *within*) pour éliminer les effets individuels α_i . Ces transformations sont-elles applicables ici ?

Question 7 Deux solutions classiques ont été proposées pour étudier les données de *panel binaire* (panel avec une variable dépendante binaire). La première consiste à supposer les effets individuels α_i indépendants des covariables et à spécifier une distribution paramétrique pour ces α_i . Le "random effect probit model" suppose $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ et est fréquemment utilisé.⁹

- Comment s'estime le "random effect probit model" ?
- Peut-on calculer les effets marginaux d'une variable explicative (ici on serait notamment intéressé par la variable "origine") dans ce modèle ?
- Estimer un "random effect probit model" avec comme précédemment pour variable dépendante l'indicatrice d'être au chômage et en choisissant et justifiant les variables explicatives utilisées. Interpréter les estimations obtenues et, selon votre réponse à la question précédente (question 7.b), calculer les effets marginaux de la variable "origine".

Question 8 Le "random effect probit model" n'autorise pas de corrélation entre les régresseurs et l'effet individuel. Le "fixed effect logit model" (parfois appelé "conditional fixed effect logit" ou encore "conditional logit estimator") relâche cette hypothèse en utilisant une transformation – similaire dans sa fonction au *first-difference* ou au *within* dans le cadre linéaire – permettant d'éliminer l'effet individuel.¹⁰

- Comment s'estime le "fixed effect logit model" ?
- Peut-on calculer les effets marginaux d'une variable explicative (ici on serait notamment intéressé par la variable "origine") dans ce modèle ?
- Estimer un "fixed effect logit model" avec comme précédemment pour variable dépendante l'indicatrice d'être au chômage et en choisissant et justifiant les variables explicatives utilisées. Interpréter les estimations obtenues et, selon votre réponse à la question précédente (question 8.b), calculer les effets marginaux de la variable "origine".

Conclusion – Question 9

- Comparer les différentes estimations réalisées sur la marge extensive (partie 2) : quelles conclusions en tirer quant à l'effet de la variable "origine" sur le fait d'être au chômage ?
- Suites à ces différentes estimations (partie 1 et partie 2), quelles sont vos conclusions quant à l'effet de la variable "origine" sur la position sur le marché du travail ? Diriez-vous qu'il existe des inégalités sur le marché du travail liées à l'origine des individus ? Les estimations réalisées suggèrent-elles des interventions publiques permettant d'agir sur ces inégalités éventuelles ?
- Dans ce projet, on a étudié séparément une marge intensive et une marge extensive. Quel autre type de modèle (vu en cours) serait-il également possible d'utiliser ici pour étudier l'effet de la variable "origine" sur le salaire parmi les actifs, à la fois actifs occupés et chômeurs ? (On ne demande pas d'estimer ce modèle).
- Quelles données supplémentaires ou autres pistes de recherche vous sembleraient être intéressantes pour approfondir votre réponse à la problématique ?

9. Pour cette question, vous pourrez consulter la section 15.8.2 du Wooldridge (*Econometrics Analysis of Cross Section and Panel Data*, 2nd edition) et utiliser la commande Stata `xtprobit`. Pour cette question sur ce nouveau modèle, comme pour la question suivante sur le "fixed effect logit model", il n'est pas attendu de longues preuves ou calculs. L'objectif principal est de découvrir et comprendre ces nouvelles méthodes.

10. Pour cette question, vous pourrez consulter la section 15.8.3 du Wooldridge (*Econometrics Analysis of Cross Section and Panel Data*, 2nd edition) et utiliser la commande Stata `xtlogit`.

Influence d'une chaîne de TV conservatrice sur le vote aux élections présidentielles américaines

Jérémy L'Hour
jeremy.l.hour@ensae.fr

March 13, 2019

Ce projet tente de quantifier l'effet de la diffusion d'une chaîne de télévision conservatrice entre 1996 et 2000 sur le vote en faveur d'un parti conservateur aux élections présidentielles américaines correspondantes. MAGA News est une chaîne d'information en continu créée en septembre 1996. Les spécificités du marché américain de la télévision câblée rendent le déploiement d'une nouvelle chaîne de TV nécessairement long et couteux : étant donnés les coûts fixes, chaque ville américaine constitue un monopole local opéré par une des quelques grandes compagnies de télévision câblée, qui font face à des contraintes sur le nombre de chaînes qu'elles peuvent diffuser. MAGA News a donc dû négocier sa diffusion auprès des compagnies de câble, souvent au détriment d'autres chaînes. Elle est considérée comme une chaîne à droite sur le spectre politique américain, dans l'absolu et relativement aux chaînes de télévision concurrentes.

Données La base de données `MAGANews.dta` regroupe des informations concernant la diffusion des chaînes de télé, les caractéristiques sociodémographiques et le vote, mesurées pour 9,265 villes américaines sur plusieurs années. Elle est disponible à l'adresse www.github.com/jlhourENSAE/MAGA-Econometrics. Dans la suite de l'énoncé, on appellera "groupe traité" l'ensemble des villes ayant accès à MAGA News en 2000 et "groupe contrôle" l'ensemble des autres villes.

Consignes importantes Pour toute réponse demandant une estimation, vous devez fournir un écart-type correspondant, dont le mode de calcul sera justifié. Une grande attention sera

portée à la façon d’exposer vos résultats. Inspirez-vous notamment de la façon dont les articles de recherche reportent les résultats de régression. Le projet est plus simple à réaliser en R ou Stata qu’en Python.

Ressources utiles Concernant les aspects statistiques de ce projet, “Large Sample Estimation and Hypothesis Testing” de Newey et McFadden (*Handbook of Econometrics*, 1994); à propos du score de propension, la partie 3 du livre *Causal Inference for Statistics, Social and Biomedical Sciences* de Imbens et Rubin (2015).

1. Statistiques Descriptives et Double-Différences

1. Quelle est la proportion des villes dans lesquelles MAGA News est diffusée en 1998, 2000 et 2003 ? En quoi cela offre-t-il la possibilité de mesurer l’impact de MAGA News sur le vote ?
2. La variable `reppresfv2p1996` (resp. `reppresfv2p2000`) mesure la part du vote conservateur aux élections présidentielles de 1996 (resp. 2000). Calculer l’estimateur des différences-de-différences de l’effet de MAGA News.
3. L’estimateur précédent est-il crédible ? Effectuer un ou plusieurs tests pour justifier votre réponse.
4. Produire des statistiques descriptives sur les caractéristiques des villes permettant d’éclairer la réponse à la question précédente.

2. Sélection et Score de Propension

Le but de cette section est d’étudier les facteurs déterminants de la présence de MAGA News dans une ville (*i.e.* la sélection dans le traitement). Pour cette partie, vous devrez considérer des variables de trois natures différentes : (1) économiques, liées au marché local de la télévision câblée, (2) liées au vote et à la couleur politique des citoyens d’une ville, (3) sociodémographiques. Plus précisément, le but de cette partie est de montrer que l’implémentation de MAGA News dans une ville n’est pas liée à deux facteurs que sont

(i) la part du vote conservateur à l'élection de 1996 (`reppresfv2p1996`) et (ii) le taux de participation à l'élection de 1996 (`totpreslvpop1996`).

1. En quoi serait-il problématique que les deux variables citées précédemment expliquent la présence de MAGA News en 2000 ?
2. Proposer trois spécifications différentes, avec un nombre croissant de variables, d'un modèle Logit où la variable expliquée est la présence de MAGA News dans une ville en 2000. La première spécification ne fera intervenir que la part du vote conservateur à l'élection de 1996 et le taux de participation à l'élection de 1996. Vous justifierez les variables de contrôle que vous ajoutez, mais pour ces variables uniquement, on ne demande pas de reporter/commenter les résultats de l'estimation. On prêtera attention, via des tests statistiques pertinents, à la qualité des modèles proposés.
3. D'après votre réponse à la question précédente, la part du vote conservateur à l'élection de 1996 et le taux de participation à l'élection de 1996 expliquent-ils l'implémentation de MAGA News ? Faire les tests correspondant.
4. Soit D_i la variable aléatoire qui vaut un si MAGA News est disponible dans la ville i en 2000 et zéro sinon, et X_i un vecteur aléatoire de dimension p mesurant des variables explicatives.

(a) On définit la variable aléatoire:

$$W_i := \frac{\mathbb{E}(1 - D_i)}{\mathbb{E}(D_i)} \exp(X_i' \beta_0).$$

Montrez que si le score de propension est donné par un modèle Logit (*i.e.* $\mathbb{P}[D = 1|X] = \exp(X' \beta_0)/(1 + \exp(X' \beta_0))$) alors:

$$\mathbb{E}[X_i | D_i = 1] = \mathbb{E}[W_i X_i | D_i = 0]. \quad (1)$$

(b) Interpréter cette équation.

5. (a) Montrer que l'équation 1 peut s'écrire comme une condition de moment du type $\mathbb{E}[g(D_i, X_i, \beta_0)] = 0$ où vous préciserez la fonction g .
Astuce: On rappelle que $\mathbb{E}[X_i|D_i = 1] = \mathbb{E}[X_i D_i] / P[D_i = 1]$.
- (b) Proposer un estimateur GMM (Méthode des Moments Généralisée) de β_0 . Le calculer pour les trois spécifications choisies à la question 2.

3. Estimation d'Impact par Régression Linéaire

Le but de cette section est d'estimer l'impact causal de la présence de MAGA News sur le vote conservateur au moyen de régressions linéaires, afin d'obtenir un estimateur meilleur que celui des différences-de-différences obtenu dans la partie 1.

*La variable **totpresvotes1996** donne le nombre de votes exprimés pour l'élection présidentielle de 1996. Dans cette section, on ponderera les résultats d'estimation par cette variable, de façon à interpréter les résultats pour l'électeur moyen, plutôt que pour la ville moyenne.*

1. En prenant en compte les résultats statistiques de la partie 2, proposez deux spécifications de la régression linéaire de la différence entre la part du vote conservateur à l'élection de 2000 et celle de 1996 (mesurée par **reppresfv2p00m96**) sur la présence de MAGA News en 2000 (mesurée par **maganews2000**). Justifiez :
 - (a) la sélection des variables de contrôle,
 - (b) la présence ou non d'une variable mesurant la différence entre la part du vote conservateur entre deux élections avant 1996,
 - (c) l'utilisation d'effets fixes,
 - (d) le calcul de l'écart-type.

Le report/commentaire des résultats d'estimation pour les coefficients associés aux variables de contrôle n'est pas demandé.

2. On note $Y_i(1)$ (resp. $Y_i(0)$) la variable aléatoire **reppresfv2p00m96** qui mesure la différence entre la part du vote conservateur à l'élection de 2000 et celle de 1996

quand MAGA News est disponible dans la ville i (resp. quand MAGA News n'est pas disponible dans la ville i). On note la variable observée $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. On définit le paramètre θ_0 :

$$\theta_0 = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[W_i Y_i | D_i = 0], \quad (2)$$

pour le W_i défini dans la partie précédente. On suppose que $Y(0) \perp\!\!\!\perp D | X$ (*Hypothèse d'Indépendance Conditionnelle*, ou CIA).

- (a) Justifier soigneusement que $\theta_0 = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$.
- (b) Proposer un estimateur de θ_0 basé sur l'équation (2), que l'on notera $\hat{\theta}$.
- (c) Montrer que $\sqrt{n}(\hat{\theta} - \theta_0)$ suit une distribution Gaussienne lorsque $n \rightarrow \infty$. Donner son écart-type asymptotique.

Astuce: On pourra utiliser, en les citant, les outils développés dans la Section 6.1 de "Large Sample Estimation and Hypothesis Testing" de Newey et McFadden (Handbook of Econometrics, 1994).

- (d) A partir de l'estimateur de β_0 calculé à la Q5 de la partie 2, calculer $\hat{\theta}$ ainsi que son écart-type. Comparer aux résultats obtenus par régression.

3. La disponibilité de MAGA News sur le réseau de télévision local impacte-t-il le vote conservateur ? Dans quelle mesure ? Proposez deux explications à ce phénomène (ou à son absence).

4. Test Placebo

On souhaite conduire un test placebo. Pour cela, on va regarder l'impact de la présence de MAGA News en 2000 sur l'évolution de la part du vote conservateur entre 1992 et 1996, ainsi qu'entre 1988 et 1992. Quel effet devrait-on observer ? Mettre en œuvre ce test pour une spécification choisie. Cela vous donne-t-il plus de confiance dans les résultats obtenus ?

Crimes et Variables Instrumentales: L'Impact de l'Arrestation sur la Probabilité de Récidive

Jérémy L'Hour
jeremy.l.hour@ensae.fr

March 13, 2019

L'utilisation d'expériences randomisées en criminologie est souvent limitée par des considérations éthiques, particulièrement lorsque la vie humaine est en jeu. Heureusement, il est toujours possible d'identifier des liens de cause à effet en utilisant des variables instrumentales. Au début des années 1980, la *Minneapolis Domestic Violence Experiment* (MDVE) a été imaginée dans le but d'évaluer l'efficacité des sanctions mises en place par les forces de police en réponse aux délits d'agression domestique. Dans cette expérience, lorsqu'un officier de police recevait un appel signalant une agression domestique répondant à des critères précis (cette expérience excluait de son champ les blessures graves), il tirait aléatoirement la réponse à adopter entre (1) l'arrestation du suspect, (2) la séparation des conjoints pour une durée de 8h ou (3) la médiation entre les conjoints. En pratique, les officiers de police déviaient souvent de l'assignation randomisée pour adopter une réponse (souvent plus sévère) mieux adaptée à la situation: typiquement, lorsque le suspect agressait l'agent de police, lorsque la victime demandait expressément à l'arrestation du suspect ou lorsque les deux parties étaient blessées, l'agent procédait à l'arrestation du suspect, peu importe l'assignation aléatoire. Il s'agit donc d'une expérience où la règle d'assignation aléatoire n'est pas toujours respectée.

Il s'agit, à travers ce projet, de quantifier à l'impact de l'arrestation d'un individu suite à un signalement pour agression domestique sur sa probabilité de récidive six mois plus tard.

Données. La base de données `MDVE.csv` regroupe des informations concernant 330 interventions de police pour agression domestique. Elle est disponible à l'adresse www.github.com/jlhourENSAE/MDVE-Econometrics.

Notations et Hypothèses. Tout au long de ce projet, on adoptera les notations suivantes. Z (`z_arrest`) est une variable aléatoire binaire qui vaut 1 si l'individu est censé être arrêté (ainsi que désigné par le tirage aléatoire) et 0 sinon. La variable aléatoire D_0 est une variable binaire qui vaut 1 si l'individu est arrêté quand l'assignation aléatoire ne prescrit pas l'arrestation ($Z = 0$) et 0 sinon. La variable aléatoire D_1 est une variable binaire qui vaut 1 si l'individu est arrêté quand l'assignation aléatoire prescrit l'arrestation ($Z = 1$) et 0 sinon. On n'observe que l'une de ces deux variables aléatoires, c'est-à-dire le traitement réellement appliqué, que l'on note $D = D_Z = D_0 + Z(D_1 - D_0)$, désignée par `d_arrest` dans les données. De façon similaire, le résultat potentiel, c'est-à-dire la récidive à six mois, est notée par Y_0 si l'individu n'a pas été arrêté ($D = 0$) et par Y_1 si l'individu a été arrêté. On n'observe que $Y = Y_0 + D(Y_1 - Y_0) = Y_0 + D_0(Y_1 - Y_0) + Z(D_1 - D_0)(Y_1 - Y_0)$, désignée par `rearrested` dans les données. On pose les hypothèses suivantes:

$$(Y_0, Y_1, D_0, D_1) \perp\!\!\!\perp Z, \quad (\text{Independance})$$

$$0 < P[Z = 1] < 1 \text{ et } P[D_1 = 1] > P[D_0 = 1], \quad (\text{First Stage})$$

$$P[D_1 \geq D_0] = 1. \quad (\text{Monotonicit })$$

On note par X un vecteur de caract ristiques du d lit. Pour chaque d lit, on observe donc le vecteur al atoire (Y, Z, D, X) et rien de plus. On supposera que les observations $(Y_i, Z_i, D_i, X_i)_{i=1, \dots, n}$ sont iid.

Consignes importantes. **Pour toute r ponse demandant une estimation, vous devez fournir un  cart-type correspondant, dont le mode de calcul sera justifi .** Une grande attention sera port e   la fa on d'exposer vos r sultats. Lorsqu'il est demand  de justifier le choix des variables de contr le, on s'attend   lire des raisonnements socio- conomiques et non   l'application d'algorithmes de Machine Learning. Inspirez-vous notamment de la fa on dont les articles de recherche reportent les r sultats de r gression. Le projet est plus simple   r aliser en R ou Stata qu'en Python.

Ressources utiles. la partie VI du livre *Causal Inference for Statistics, Social and Biomedical Sciences* de Imbens et Rubin (2015) ou la partie 4 du livre *Mostly Harmless Econometrics* d'Angrist et Pischke (2008).

1. Statistiques Descriptives

Produire quelques statistiques descriptives pertinentes par rapport au problème soulevé par le sujet. Les commenter de manière synthétique.

2. Modèle Logit Simple

On s'intéresse au coefficient τ_0 donnant l'impact causal de l'arrestation sur la récidive à six mois, c'est à dire:

$$\tau_0 = E_X [P[Y = 1|X, D = 1] - P[Y = 1|X, D = 0]],$$

avec $E_X[.]$ l'espérance prise en X . Pour cela, on utilise la modélisation suivante:

$$Y = \mathbf{1}\{\alpha_0 + \gamma_0 D + X'\beta_0 - U \geq 0\}, \quad (1)$$

où X est un vecteur de variables de contrôle et U est un terme d'erreur de fonction de répartition $F_U(t) = 1/(1 + \exp(-t))$.

1. En supposant que $U \perp\!\!\!\perp (D, X)$, calculer l'estimateur du maximum de vraisemblance de $(\alpha_0, \gamma_0, \beta'_0)'$ à partir des données dans un cas sans variable de contrôle ($\beta_0 = 0$) et dans un autre cas où vous utiliserez des variables de contrôle (X) dont vous justifierez le choix. On écrira la vraisemblance de façon détaillée. Faites les tests nécessaires et commentez au moyen des outils introduits dans le cours pour ce genre de modèle.
2. Donner un estimateur de τ_0 . Calculer et commenter.
3. Selon vous, pourquoi il est peu crédible de supposer $U \perp\!\!\!\perp (D, X)$? Que cela signifie-t-il pour les deux questions précédentes?
4. Dites pourquoi la variable Z peut être un bon instrument pour D , l'arrestation du suspect.

3. Etude de la réponse à l'assignation aléatoire (First Stage)

A partir du couple de variables aléatoires (D_0, D_1) , on considère quatre sous-populations: les *always-takers* pour lesquels $D_1 = D_0 = 1$, les *never-takers* pour lesquels $D_1 = D_0 = 0$, les *compliers* pour lesquels $D_1 > D_0$ et les *defiers* pour lesquels $D_1 < D_0$.

1. D'après les hypothèses (Independance), (First Stage) et (Monotonicité) introduites plus haut, laquelle de ces quatre sous-populations n'existe pas?
2. Dans le contexte de la MDVE, que signifie l'événement $\{D = 1|Z = 0\}$? Montrer que $P[D = 1|Z = 0] = P[D_0 = D_1 = 1]$. Peut-on identifier la proportion d'*always-takers* dans la population? Si oui, proposer un estimateur convergent de cette quantité et le calculer à partir des données.
3. Dans le contexte de la MDVE, que signifie l'événement $\{D = 0|Z = 1\}$? Peut-on identifier la proportion de *never-takers*? Si oui, proposer un estimateur et le calculer à partir des données. Doit-on s'attendre, dans le contexte de l'expérience, à ce que cette proportion soit grande? Pourquoi?
4. Comment qualifieriez vous un individu tel que $D_1 > D_0$? A quelle situation d'agression cela correspond-il probablement?
5. Estimer deux modèles de régression linéaire de D sur Z : un sans variables de contrôle, puis un autre avec les variables de contrôle qui vous semblent pertinentes (vous justifierez vos choix). Tester la condition de rang et conclure.

4. Stratégie 2SLS

On propose donc d'instrumenter D par Z dans le modèle linéaire

$$Y = \tilde{\alpha}_0 + \tilde{\tau}_0 D + X' \tilde{\beta}_0 + \varepsilon, \text{ avec } E[\varepsilon|X, Z] = 0.$$

1. Donner la limite en probabilité de l'estimateur des variables instrumentales dans le modèle sans variable de contrôle, $Y = \tilde{\alpha}_0 + \tilde{\tau}_0 D + \varepsilon$ avec $E[\varepsilon|Z] = 0$. Montrer

soigneusement que cette limite est égale à $E[Y_1 - Y_0 | D_1 > D_0]$. Interpréter cette quantité.

2. Etant donnée votre réponse à la question (3) de la partie 3 et l'expérience mise en place, pouvez-vous justifier l'hypothèse selon laquelle $P[D_1 = 1] = 1$? Qu'est ce que cela implique pour vos quatre populations sous-jacentes? Montrer que dans ce cas $E[Y_1 - Y_0 | D_1 > D_0] = E[Y_1 - Y_0 | D = 0]$. Comment interpréter cette quantité?
3. Estimer $\tilde{\tau}_0$ sans et avec les variables de contrôle de votre choix. On détaillera la méthode employée dans les deux cas, ainsi que les choix des variables de contrôle.
4. Effectuer les tests qui vous semblent appropriés étant donné le contexte.
5. Selon vous, l'arrestation du suspect lorsque la police est appelée pour une affaire d'agression domestique est-elle une solution efficace pour lutter contre la récidive? Si oui, sur quel type d'individus? Que dire des autres?

5. Analyse des Compliers via le Kappa d'Abadie (2003)

1. Pourquoi est-il difficile d'étudier les individus tels que $D_1 > D_0$?
2. Montrer soigneusement que $P[D_1 > D_0] = E[D | Z = 1] - E[D | Z = 0]$. Comment pourrait-on estimer $P[D_1 > D_0]$ de façon convergente? Le faire à partir des données.
3. On adopte les deux hypothèses suivantes:

$$(Y_0, Y_1, D_0, D_1) \perp\!\!\!\perp Z | X, \quad (\text{Independance'})$$

$$P[Z = 1 | X] = P[Z = 1] = \pi_0, \quad (\text{Randomization})$$

et on note κ la variable aléatoire suivante:

$$\kappa = \frac{1}{P[D_1 > D_0]} \frac{D(Z - \pi_0)}{\pi_0(1 - \pi_0)}.$$

Montrer que $E[X | D_1 > D_0] = E[\kappa X]$ sous ces hypothèses.

Astuce: On pourra s'aider de l'article "Semiparametric instrumental variable estimation of treatment response models" d'Abadie (2003).

4. Proposer une stratégie pour estimer de façon convergente les caractéristiques des compliers, *i.e.* $E[X|D_1 > D_0]$. On présentera l'estimateur sous la forme d'un estimateur GMM.
5. Comparer les compliers au reste de la population à partir de vos données. Commenter.

Projet d'Econométrie 2 :

La beauté et la couleur de peau en politique

Chargée de TD :

Ophélie Couperier (o.couperier@gmail.com)

Données disponibles sur :

<https://ocouperier.wixsite.com/website>

Une caractéristique inhabituelle des élections dans le Territoire du Nord, en Australie, est que les photographies des candidats apparaissent sur le bulletin de vote à côté de leur nom (voir exemple en fin d'énoncé). En vertu du règlement électoral, tous les candidats sont tenus de présenter un portrait vertical en noir et blanc de la tête et des épaules du candidat, dans les six mois précédant le dépôt de la candidature. Les élections de 2005, sur lesquelles se concentrent ce projet, n'étaient pas les premières élections dans le Territoire du Nord où figuraient des photographies des candidats, mais c'était la première élection dans le Territoire du Nord où les bulletins de vote montraient à la fois des photographies et le nom des partis. L'objectif de ce projet est de mieux comprendre dans quelle mesure la couleur de peau et la beauté du candidat sont liés au comportement des électeurs.

Les données : Les données de vote proviennent du site Web du Bureau électoral du Territoire du Nord, qui publie le nombre de votes reçus par chaque candidat et les profils démographiques de chaque électorat. Elles portent sur les élections du 18 juin 2005¹ dans le Territoire du Nord. La base de données fournit des informations sur plusieurs caractéristiques de l'électorat, notamment la part de la population qui s'identifie comme autochtone ou insulaire du détroit de Torres.

La base de données contient également des informations sur les candidats : le sexe, le parti politique, le statut de président sortant, la beauté et la couleur de peau. Pour la beauté des candidats, les photographies du bulletin de vote des 79 candidats² ont été regroupées dans un document PDF qui a ensuite été présenté à des « noteurs » qui devaient évaluer la beauté de chaque visage sur une échelle de 1 à 9 (1 étant le moins attrayant et 9 le plus attrayant). Les noteurs ont été choisis de sorte à considérer à la fois la population autochtone et non autochtone dans le Territoire du Nord, et représenter le sexe et l'âge de l'électorat australien. Ainsi les quatre voteurs présentaient les caractéristiques suivantes : un homme autochtone de 24 ans, une femme autochtone de 39 ans, une femme non autochtone de 24 ans, un homme non autochtone de 40 ans. Pour la couleur de peau des candidats, une échelle de nuancier allant de blanc

¹ Le Parlement du Nord dispose d'un parlement monocaméral et de 25 sièges à l'Assemblée législative. Lors des élections précédentes (tenues en 2001), le parti travailliste australien (Australian Labor Party) de centre-gauche avait été élu pour la première fois depuis que le Territoire du Nord avait atteint l'autonomie gouvernementale en 1978. Lors des élections de 2005, le parti travailliste australien a encore augmenté sa majorité avec 19 des 25 sièges.

² Il y avait 80 candidats mais l'un d'entre eux a refusé de faire partie de l'étude.

(noté 0) à noir (noté 16) a été utilisée puisque les photos fournies sur le bulletin de vote sont en noir et blanc (voir l'échelle utilisée en fin d'énoncé).

Les consignes : Les résultats des estimations effectuées doivent apparaître dans le rendu du projet (estimations, écart-types). Vous devez interpréter vos résultats. Une attention particulière sera accordée à la qualité de la rédaction (interprétations des résultats) et à la justification des modélisations proposées (Quelles sont les hypothèses des modèles ? Sont-elles valides ?). Justifiez le choix des variables (choix des variables de contrôle, introduction d'effet fixe, de variables croisées, ...). Prenez également soin de rappeler les hypothèses nulle et alternative des tests statistiques que vous effectuerez.

Questions :

Question 1 :

- (i) Quelle est la particularité principale du vote en Australie ? Quelles sont les sanctions encourues ? Le taux de participation est-il élevé pour autant ? Vous pouvez vous aider du papier de Shephard (2005) pour répondre à ces questions.
- (ii) A priori, pensez-vous que la beauté et/ou la couleur de peau influent les électeurs dans leur choix de vote ? Argumentez votre réponse.

Question 2 : Etablissez le tableau des corrélations entre les évaluateurs. Interprétez. Ces résultats sont-ils semblables à ceux attendus dans la littérature ? Vous pouvez vous aider des articles de Langlois et al (2000) et Rhodes (2006) pour répondre à la question.

Question 3 : Donnez les moyennes et variances des scores de beauté des quatre noteurs et interprétez. Redimensionnez les scores de chaque évaluateur en un z-score (moyenne 0 et écart-type 1). Créez une nouvelle variable synthétique des scores de beauté et justifiez votre choix.

Question 4 : Donnez les statistiques descriptives des variables pertinentes et interprétez.

Question 5 :

- (i) Graphique pour chaque candidat la part de vote obtenue en fonction du score de beauté. Ajustez une droite de régression sur le graphique. Ecrivez l'équation générale du modèle. Quelles hypothèses supposez-vous pour appliquer le modèle ?
- (ii) Même question en séparant les candidats sortants des opposants. Interprétez les résultats.
- (iii) Mêmes questions en considérant la part de vote obtenue en fonction de la couleur de peau (sur l'échantillon entier et en dissociant candidats sortants et opposants). Les résultats sont-ils similaires ? Pourquoi ?

- (iv) Constituez deux sous-groupes de tailles similaires réalisés à partir de la part d'individus autochtone dans l'électorat (variable ab). Faites à nouveau le point (iii) en considérant ces sous-groupes (vous obtenez 4 nouveaux graphiques). Interprétez.

Question 6 :

Supposons que l'électeur i ait une utilité associée à l'élection du candidat j telle que : $U_{ij} = X_j' \beta + \xi_j + \varepsilon_{ij}$ où X_j correspond aux caractéristiques observées du candidat j et ξ_j correspond aux caractéristiques inobservées. L'électeur peut également choisir de voter blanc (choix $j = 0$), l'utilité correspondante étant alors $U_{i0} = \varepsilon_{i0}$. On suppose que les $(\varepsilon_{ij})_{j=0,\dots,J}$ sont i.i.d. de loi de Gompertz. L'électeur i vote pour le candidat j qui maximise son utilité d'où le choix de vote Y_i tel que : $Y_i = \arg \max_{j=0,\dots,J} U_{ij}$.

- (i) Calculer $Pr[Y = j | X_1, \dots, X_J]$. En déduire que : $\ln Pr[Y = j | X_1, \dots, X_J] - \ln Pr[Y = 0 | X_1, \dots, X_J] = X_j' \beta + \xi_j$ pour tout $j \neq 0$.
- (ii) Comment se modifie le ratio entre la part des votes allant vers deux candidats lorsqu'un nouveau candidat se présente ? Comment se nomme cette propriété du modèle ?
- (iii) Pensez-vous, étant donné le contexte, que cette hypothèse soit raisonnable ?
- (iv) Dans la base de données, nous observons seulement les parts de marchés (i.e. le pourcentage de votes exprimés), proposez une méthode d'estimation de β convergente, sous des hypothèses que vous préciserez.

Question 7 :

- (i) La beauté et la couleur de peau ont-elles un impact sur les parts de vote obtenues par les candidats ? Proposez six modélisations différentes. Vous utiliserez notamment votre réponse à la question 6, iv.
- (ii) D'après votre réponse à (i), la beauté et la couleur de peau du candidat expliquent-elles la part des votes obtenue par les candidats ? Justifiez votre réponse par des tests adaptés.

Question 8 :

- (i) L'hypothèse d'exogénéité de la variable de beauté nécessaire à la validité des modèles linéaires utilisés à la question précédente vous semble-t-elle vérifiée ? Pourquoi ?
- (ii) Proposez un instrument que vous justifierez soigneusement.
- (iii) Ecrivez les modèles de régression de première étape. Estimez les modèles et discutez de la qualité de l'instrument. On pensera à faire les tests appropriés.
- (iv) Interprétez les résultats des estimations finales. Les conclusions sont-elles identiques à celles de la question 7 ?

Question 9 : Vous avez à disposition dans la base de données la variable *winner* qui prend la valeur 1 si le candidat est finalement élu, 0 sinon.

- (i) Proposez six spécifications différentes pour modéliser cette variable.
- (ii) D'après votre réponse à (i), la beauté et la couleur de peau influent-elles sur la probabilité d'être élu ? Justifiez par des tests.

Question 10: Vous allez maintenant considérer un modèle probit avec variable instrumentale. Vous considérerez la méthode de Rivers et Vuong (1989) vue lors du TD 6-7 exercice 2.

- (i) Ecrivez les modèles. Quelles hypothèses doivent être satisfaites pour la mise en œuvre ?
- (ii) Effectuez les estimations avec cette méthode. Commentez et comparez aux résultats obtenus à la question précédente.

Question 11 : La variable *winner2* représente le nombre de fois où le candidat a été élu lors des deux dernières élections.

- (i) Quel est le modèle à considérer ? Donnez l'écriture du modèle. Énoncez et discutez les hypothèses de ce modèle. Comment l'estimez-vous ?
- (ii) Estimez le modèle et commentez les résultats. La beauté et la couleur de peau ont-elles un effet sur le nombre d'élections du candidat ?
- (iii) Donnez l'écriture de $Pr[winner2 \leq i|X]$ pour $i \in \{0; 1; 2\}$. Quelle hypothèse qui peut paraître restrictive est faite ici ? Effectuez le test de pentes parallèles. En cas de rejet, quel modèle est à privilégier ? Justifiez. D'après le test ce modèle est-il valide ?

Références

- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). *Maxims or myths of beauty? A meta-analytic and theoretical review*. *Psychological Bulletin*, 126(3), 390-423.
- Rhodes, G. (2006). *The Evolutionary Psychology of Facial Beauty*. *Annual Review of Psychology*, 57, 199-226.
- Shephard, B. (2005). *Northern Territory Electoral Commission Annual Report 2004-2005*. Darwin: NTEO.

Exemple de bulletin de vote avec photographies :

Form 1
Regulation 4




NORTHERN TERRITORY OF AUSTRALIA
Electoral Act

BALLOT PAPER

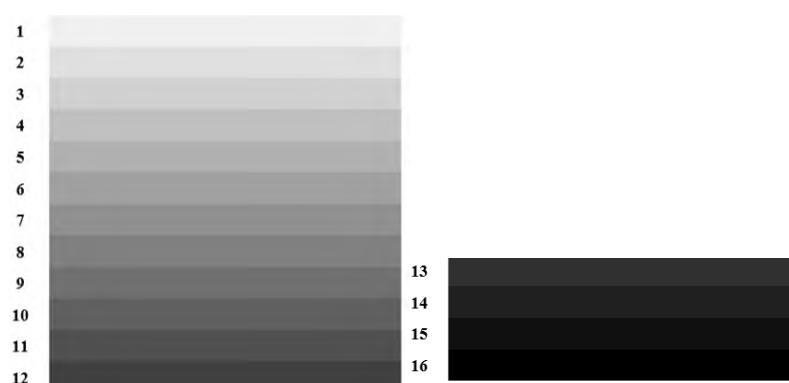
ELECTION OF ONE MEMBER OF
THE LEGISLATIVE ASSEMBLY FOR THE
DIVISION OF ARAFURA

Directions:
Number the boxes 1 to 3 in the order of your choice. Remember, number every
box to make your vote count.

CANDIDATES

<input type="checkbox"/>		SCRYMGOUR, Marion Australian Labor Party NT (ALP)
<input type="checkbox"/>		STEVENS, August Northern Territory Country Liberal Party
<input type="checkbox"/>		PASCOE, George The Greens

Echelle utilisée pour donner la modalité à la variable de couleur de peau :



Impact du salaire minimum dans la restauration rapide: une réflexion autour de l'article de Card et Krueger (1994)

Yannick Guyonvarch

1 Présentation du problème et de la base de données

Ce projet est basé sur les données de l'article de D. Card et A.B. Krueger datant de 1994 et intitulé "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". Dans cet article, les deux auteurs étudient l'impact d'une hausse du salaire minimum légal sur l'emploi dans le secteur de la restauration rapide. Pour ce faire, D. Card et A.B. Krueger comparent la situation des établissements de restauration rapide dans deux états limitrophes aux Etats-Unis (le New Jersey et la Pennsylvanie) et utilisent le fait que seul le New Jersey introduit une hausse du salaire minimum légal. Leur cadre d'analyse correspond donc à celui d'une différence-de-différence (DID) classique. Par la suite, nous noterons Y la variable capturant l'emploi, G une indicatrice de localisation dans le New Jersey, T une indicatrice valant 0 si les établissements de restauration rapide sont observés avant la réforme et 1 s'ils sont observés après et $D = G \times T$ est la variable dite de traitement. Par la suite, nous supposons avoir un échantillon i.i.d de n restaurants pour lesquels nous observons (Y, G, T) . Enfin pour tout $(i, j) \in \{0, 1\}^2$, nous notons $\mathbb{E}[Y_{ij}] = \mathbb{E}[Y \mid G = i, T = j]$.

2 Statistiques descriptives

Comparer les différentes sous-populations de l'échantillon (restaurants du New Jersey *vs* en Pennsylvanie, restaurants avant et après la réforme...) en utilisant les outils statistiques qui vous semblent pertinents (comparaisons à la moyenne, à différents percentiles...).

3 DID

Nous nous plaçons dans un cadre "à la Rubin" (1974): nous supposons observer $Y = DY(1) + (1-D)Y(0)$ avec Y et D définis comme précédemment. $Y(0)$ et $Y(1)$ sont des niveaux d'emploi potentiels et nous n'observons que l'un des deux selon le statut de traitement. Notre but dans cette section est d'estimer le paramètre $\Delta = \mathbb{E}[Y(1) | G = 1, T = 1] - \mathbb{E}[Y(0) | G = 1, T = 1]$.

Question 3.1: $\mathbb{E}[Y(1) | G = 1, T = 1]$ et $\mathbb{E}[Y(0) | G = 1, T = 1]$ sont-elles directement estimables dans les données? Soit $DID := \mathbb{E}[Y_{11}] - \mathbb{E}[Y_{10}] - \mathbb{E}[Y_{01}] + \mathbb{E}[Y_{00}]$. Sous quelles conditions a-t-on $\Delta = DID$? Comment peut-on estimer Δ à l'aide de cette relation? Montrer la normalité asymptotique de l'estimateur correspondant et expliciter sa variance limite.

Question 3.2: Posons le modèle $Y = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 D + \epsilon$. Sous quelles conditions sur (ϵ, G, T) a-t-on $\beta_3 = \Delta$? Comment peut-on estimer Δ ici? Rappeler la loi limite de l'estimateur (pas besoin de la redémontrer).

Question 3.3: Estimer Δ par l'une des deux méthodes précédentes. Pour ce faire, plusieurs variables dans la base de données fournissent une information sur l'emploi. Justifier quelle(s) variable(s) vous pensez être la/les plus pertinente(s). Vous pouvez également vous inspirer de Card and Krueger (1994) pour construire votre propre variable. Mener le test d'hypothèse de nullité de Δ . Ne pas oublier de rappeler la statistique de test, les hypothèses nulle et alternative, la loi asymptotique de la statistique sous les différentes hypothèses.

Question 3.4: Nous avons à disposition des variables explicatives supplémentaires que nous souhaitons mettre à profit dans l'estimation de Δ . Nous notons $X \in \mathbb{R}^p$ ces covariables. Pour tout $(i, j) \in \{0, 1\}^2$, nous notons $\mathbb{E}[Y_{ij} | X] = \mathbb{E}[Y | G = i, T = j, X]$ et nous définissons la quantité suivante

$$DID_X := \mathbb{E}[Y_{11}] - \mathbb{E}\left[\mathbb{E}[Y_{10} | X] + \mathbb{E}[Y_{01} | X] - \mathbb{E}[Y_{00} | X] \mid G = 1, T = 1\right].$$

Sous quelles conditions sur $(Y(0), G, T, X)$ avons-nous $\Delta = DID_X$? Quel est l'avantage des hypothèses qui permettent d'obtenir $\Delta = DID_X$ par rapport à celles qui impliquent $\Delta = DID$?

Question 3.5: En supposant que pour tout $(i, j) \in \{0, 1\}^2$, $\mathbb{E}[Y_{i,j} \mid X] = \beta_{0,(i,j)} + X' \beta_{1,(i,j)}$, comment estimeriez-vous $(\beta_{0,(i,j)}, \beta'_{1,(i,j)})'$? En déduire un estimateur de $x \mapsto \mathbb{E}[Y_{i,j} \mid X = x]$ et enfin un estimateur de Δ . Dans cette question, il est seulement demandé de donner la formule de l'estimateur.

Question 3.6: Nous admettons que l'estimateur précédent est asymptotiquement normal et que la méthode dite du Efron's percentile bootstrap (Efron (1979)) est valide pour construire un intervalle de confiance pour Δ . Cette méthode est décrite dans l'Algorithme 1. Estimer Δ en incorporant les covariables qui vous paraissent intéressantes (une justification du choix des variables explicatives est attendu). Utiliser la méthode du Efron's percentile bootstrap pour construire un intervalle de confiance au niveau de votre choix pour Δ (il faut utiliser un grand nombre de réplifications bootstrap, *i.e* au moins 200).

Algorithm 1: Intervalle de confiance (IC) basée sur l'*Efron's percentile bootstrap*

Input: Les données $V_i = (X_i, Y_i, G_i, T_i), i = 1, \dots, n, n > 0$.

Input: Un niveau de confiance $1 - \alpha$

Input: Un nombre de réplifications bootstrap B

for $b = 1, \dots, B$ **do**

Tirer n fois avec remise dans $(V_i)_{i=1}^n$ pour obtenir un échantillon bootstrap de taille

n noté $(V_i^{(b)})_{i=1}^n$;

Calculer $\Delta^{(b)}$, l'estimateur obtenu sur le b -ème échantillon bootstrap ;

end

Calculer $\Delta_{\alpha/2}^B := \inf_{s \in \{1, \dots, B\}} \left\{ \Delta^{(s)} : \frac{1}{B} \sum_{b=1}^B \mathbb{1} \{ \Delta^{(b)} \leq \Delta^{(s)} \} \geq \frac{\alpha}{2} \right\}$;

Calculer $\Delta_{1-\alpha/2}^B := \inf_{s \in \{1, \dots, B\}} \left\{ \Delta^{(s)} : \frac{1}{B} \sum_{b=1}^B \mathbb{1} \{ \Delta^{(b)} \leq \Delta^{(s)} \} \geq 1 - \frac{\alpha}{2} \right\}$;

Output: Un IC de niveau asymptotique $1 - \alpha$ pour Δ , noté $I_{n,B}^{1-\alpha} := [\Delta_{\alpha/2}^B, \Delta_{1-\alpha/2}^B]$.

4 Déterminants du niveau des salaires

Dans cette section, nous ne nous intéressons qu'à la période qui précède la hausse du salaire minimum. La variable d'intérêt est $Y = \mathbb{1} \{ \text{les salaires sont élevés} \}$. Nous supposons que le modèle suivant est valide $Y = \mathbb{1} \{ X' \beta + \epsilon > 0 \}$ avec $\epsilon \mid X \sim \text{Logistique}$ et X un ensemble de variables explicatives.

Questionb 4.1: Construire (en justifiant) la variable Y qui vous paraît appropriée. Justifier soigneusement le choix des variables explicatives.

Questionb 4.2: Estimer β par maximum de vraisemblance en utilisant toutes les observations avant le traitement. Ecrire la vraisemblance du modèle et rappeler les conditions sous lesquelles l'estimateur obtenu est asymptotiquement normal. Estimer les effets marginaux moyens associés aux différentes variables explicatives (en rappelant leur formule bien sûr). Tester la nullité des coefficients de β et des effets marginaux.

Question 4.3: Nous voulons désormais tester l'hypothèse que le modèle Logit est le même dans le New Jersey et en Pennsylvanie, avant le traitement. Autrement dit, nous supposons que le modèle s'écrit $Y = \mathbb{1} \{X' \beta_g + \epsilon > 0\}$ avec $\epsilon | X \sim \text{Logistique}$ et g une variable muette qui indique l'appartenance à chaque état (0 pour la Pennsylvanie, 1 sinon). L'hypothèse que l'on souhaite tester est la suivante $\beta_0 = \beta_1$. Effectuer un tel test en adaptant le test de Chow présenté habituellement dans les modèles linéaires (il vous faudra rappeler la statistique de test, sa loi asymptotique ainsi que l'hypothèse nulle que vous testez).

Question 4.4: Nous voulons répondre à une question proche de celle de 4.3 mais avec une méthode alternative. Le paramètre que l'on cherche à estimer est $\theta := \mathbb{E} [\Lambda(X' \beta_0) | G = 0, T = 0] - \mathbb{E} [\Lambda(X' \beta_1) | G = 0, T = 0]$ où $\Lambda(\cdot)$ est la fonction de répartition de la loi Logistique.

Question 4.4.1: Proposer un estimateur de θ basé sur β_{n0} et β_{n1} les estimateurs du maximum de vraisemblance dans les sous-populations $\{G = 0, T = 0\}$ et $\{G = 1, T = 0\}$. Nous notons θ_n cet estimateur.

Question 4.4.2: En admettant la normalité asymptotique de $\sqrt{n}(\theta_n - \theta)$ et la validité de l'Efron's percentile bootstrap, implémenter un intervalle de confiance sur θ utilisant l'Efron's percentile bootstrap (au niveau de confiance de votre choix).

Question 4.4.3: Si l'intervalle de confiance ne contient pas 0, que pouvez-vous en déduire sur β_0 et β_1 ?

References

- CARD, D. AND A. B. KRUEGER (1994): “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,” *American Economic Review*, 84, 772–793.
- EFRON, B. (1979): “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, 7, 1–26.

Estimation des rendements privés et sociaux de l'enseignement supérieur

Elio Nimier-David *

1 Introduction

Depuis le début des années 1970, la France connaît une forte expansion du nombre de diplômés du supérieur. Entre 1970 et 2010, le nombre d'étudiants inscrits dans l'enseignement supérieur a été multiplié par 2,7 selon les chiffres de l'INSEE. Ce fort développement a notamment été rendu possible par la création de nombreux établissements sur l'ensemble du territoire.

Ce projet propose de mettre à profit cette "massification" de l'enseignement supérieur afin d'estimer les rendements sociaux et privés de l'éducation. Ainsi, on cherchera dans un premier temps à estimer l'effet d'être diplômé du supérieur sur le niveau de salaire de l'individu. On cherchera ensuite à voir si l'augmentation généralisée du niveau de diplôme (i.e. l'augmentation de la part de la population diplômée du supérieur) a un effet au niveau individuel sur le salaire, à niveau de diplôme donné. Cela revient à se demander si les salariés sont plus productifs, et donc mieux payés, lorsqu'ils sont entourés de travailleurs plus qualifiés (toutes choses étant égales par ailleurs).

Pour mener à bien ce projet, nous disposons de données d'enquête au niveau individuel pour la période 2003-2014. Ces données de panel nous permettent de suivre les salariés français sur plusieurs années. Elles nous renseignent sur leur salaire, leur niveau de diplôme, ainsi que les principales variables sociodémographiques (sexe, âge, situation familiale, localisation, etc.). Ces données sont ensuite complétées par le répertoire national des établissements qui recense la date et le lieu d'ouverture de nouveaux établissements de l'enseignement supérieur.¹

Remarque: ces données vous sont rendues accessibles dans le cadre de ce projet. **Elles ne doivent être ni diffusées ni exploitées dans le cadre d'autres projets.**

Pour chaque modèle estimé, une attention particulière sera portée à :

- la méthode d'estimation
- la spécification adoptée (transformations de la variable dépendante, choix des variables de contrôle, etc.)
- l'interprétation des résultats obtenus (coefficients, écarts-types et significativité)
- la réalisation des tests pertinents
- l'interprétation et la justification économique des résultats obtenus
- la précision de la rédaction

*elio.nimier-david@ensae.fr

¹Le détail des variables et la base de données seront transmis ultérieurement aux élèves travaillant sur ce projet.

Il est attendu que vous ayez un regard critique sur les spécifications employées, les hypothèses sous-jacentes et les résultats obtenus. Vous veillerez à proposer pour chaque modèle les écarts-types les plus adaptés.

Remarque: Vous devrez impérativement commenter votre code.

2 Rendements privés de l'éducation et équation de Mincer

1. Décrivez précisément deux mécanismes économiques expliquant l'impact de l'éducation sur le salaire. La référence à des théories économiques standards sera valorisée.
2. Présentez quelques statistiques descriptives permettant de rendre compte des principales variables d'intérêt. Y a-t-il des observations "aberrantes" ? Pourquoi s'intéresse-t-on aux outliers et que peut-on faire dans ce cas ?
3. Tracez l'évolution du niveau de salaire, du niveau d'éducation ainsi que le "college premium"² sur la période considérée. Vous tracerez sur le même graphique le "College Premium" pour les salariés ayant une licence, ceux ayant un master ou plus, et ceux ayant étudié dans le supérieur sans obtenir l'un de ces diplômes. Commentez. Vous pourrez présenter des résultats en fonction du sexe et de l'âge lorsque cela vous semblera pertinent.

On négligera pour la prochaine question l'aspect panel des données. On considèrera que chaque observation représente un individu.

4. Afin d'évaluer l'effet de l'éducation sur le salaire, estimez le modèle suivant :

$$\ln(w_{it}) = \beta_0 + \beta_1 s_{it} + \beta_2 x_{it} + \beta_3 x_{it}^2 + \epsilon_{it}$$

avec w_{it} le salaire mensuel de l'individu "i" l'année "t", s_{it} le nombre d'années d'études, et x_{it} l'expérience potentielle³.

Interprétez les résultats obtenus. Sous quelles hypothèses les coefficients sont-ils estimés sans biais ? Discutez la validité de chaque hypothèse et proposez des corrections si besoin.

5. Les données existent sous forme de panel. Peut-on exploiter cette caractéristique afin d'estimer sans biais l'effet de l'éducation sur le salaire, ou au moins de faire face à certaines critiques du modèle précédent ? Expliquez.
6. Certaines personnes refusent de donner précisément leur niveau de salaire. Dans ce cas, l'INSEE demande dans quelle tranche de revenus l'individu se situe⁴. On suppose dans cette question qu'on ne dispose que d'une variable de salaire par "tranche". Proposez et estimez un ou plusieurs modèles adaptés. Qu'en concluez-vous ? Vous omettez pour cette question l'aspect panel des données.

3 Les rendements sociaux de l'éducation

Dans cette deuxième partie, on cherche à estimer les rendements sociaux de l'éducation, définis comme l'impact du niveau général d'éducation sur le salaire individuel. Plusieurs travaux montrent que l'éducation produit des externalités qui affectent la productivité des autres travailleurs et donc leur salaire. On souhaite par conséquent inclure cet effet indirect de l'éducation dans notre modèle. En règle générale, on se restreint

²On définit le "College Premium" comme le ratio entre le salaire moyen des individus diplômés de l'enseignement supérieur et le salaire moyen des individus ayant uniquement le baccalauréat.

³Il s'agit de l'équation de Mincer (1974), l'un des modèles les plus connus dans le champ

⁴Les bornes des tranches sont fixées au préalable par l'INSEE.

à l'échelle de la ville dans laquelle l'individu travaille. Cette information n'étant pas disponible dans nos données, on se placera à l'échelle du département.

7. On propose d'estimer la spécification suivante :

$$\ln(w_{it}) = \alpha + \beta s_{it}^{ind} + \delta S^{dep_{it}} + \gamma X_{it}^{ind} + \rho X^{dep_{it}} + \eta_{it}$$

avec s_{it}^{ind} le niveau de diplôme de l'individu⁵, $S^{dep_{it}}$ le pourcentage de salariés diplômés du supérieur dans le département où travaille l'individu "i" à la date "t", X_{it}^{ind} des caractéristiques socio-démographiques individuelles, et $X^{dep_{it}}$ les caractéristiques du département. Estimez le modèle et commentez les résultats.

8. Quelles sont les limites de ce modèle ?

Dans la suite de cette section, on cherchera à estimer les rendements sociaux de l'éducation en utilisant la méthode des variables instrumentales.

9. Estimez le modèle précédent en différence première et en instrumentant la variable pertinente. Vous pourrez proposer un ou plusieurs instruments en discutant précisément leur validité. Un exemple d'instrument pourrait être les créations d'établissements d'enseignement supérieur. Vous utiliserez l'instrument ou les instruments les mieux adaptés.

10. Pour estimer simultanément les rendements privés et les rendements sociaux de l'éducation, il faut instrumenter le niveau de diplôme individuel et général. Quels instruments pourrait-on utiliser à cet usage ? Vous pourrez vous inspirer de la littérature sur le sujet. Commentez les résultats obtenus.

4 Conclusion

11. Synthétisez et discutez les principaux résultats obtenus. Quelles limites voyez-vous dans l'analyse menée ? Quelles pistes vous semblent intéressantes pour approfondir ce sujet ?

5 Bibliographie

Glaeser, Edward L., Ming, Lu (2018). "Human-capital externalities in China", NBER working paper.

Moretti, Enrico (2003). "Human capital externalities in cities", NBER working paper.

Moretti, Enrico (2004). "Estimating the social return to higher education, evidence from longitudinal and repeated cross-sectional data", *Journal of Econometrics*. 121: 175-212.

Mincer, Jacob (1958). "Investment in Human Capital and Personal Income Distribution". *Journal of Political Economy*. 66 (4): 281-302.

Mincer, J. (1974). "Schooling, Experience and Earnings". New York: National Bureau of Economic Research.

⁵Vous diviserez la population en 4 groupes: sans diplôme, baccalauréat, licence, master et plus.

Indice immobilier et biais de sélection

Roxane Morel

Mars 2019

Ce projet vise à créer un indice de prix de transaction pour les ventes d'appartements anciens à Paris.

Contexte Les difficultés à se loger en région parisienne sont un problème que les différentes politiques publiques des gouvernements successifs n'ont pas su résoudre. Si l'on en croit les statistiques publiques sur l'immobilier en France, les prix auraient triplé depuis les années 90. En cause : le développement du marché financier et l'allongement des durées de crédit. La tendance semble se poursuivre actuellement, notamment grâce aux taux d'intérêt très bas. Mais comment mesure-t-on cette hausse ?

En effet, le marché immobilier est un marché particulièrement imparfait. Tous les biens sont hétérogènes par définition, puisqu'on ne peut pas avoir deux appartements exactement au même endroit (ils seront au mieux voisins de palier ou à un étage différent), il ne suffit donc pas de comparer un prix médian par période : il y a fort à parier que les types de biens vendus varient au cours du temps. Par ailleurs, rien ne dit que les transactions sont représentatives du parc immobilier : les appartements familiaux sont peut-être vendus moins souvent que les studios par exemple. La méthodologie de l'indice mérite donc qu'on s'y attarde.

Données Les indices immobiliers produit par l'Insee se basent sur les données des notaires (en effet, chaque vente dans l'ancien doit passer devant un notaire, qui est tenu d'alimenter une base nationale). Cependant, une autre source existe depuis peu : DV3F (demande de valeur foncières et fichiers fonciers). C'est une base créée par l'administration fiscale, qui enregistre elle aussi les ventes, et avec une meilleure exhaustivité. La base dont vous disposerez aura été anonymisée et bruitée; elle ne pourra en aucun cas être utilisée dans un autre cadre et devra être détruite lors du rendu du projet.

1 Statistiques descriptives

1. Calculer les statistiques usuelles (min, max, écart-type, moyenne...) pour les variables numériques et réaliser des tableaux de fréquence pour les variables catégorielles de la table "vente".
2. Réaliser une matrice de corrélation des variables pertinentes de la table "ventes".

3. Comparer les caractéristiques des biens vendus et des propriétaires à celles du stock. Les ventes sont-elles représentatives de l'ensemble de l'immobilier parisien ? Justifier la réponse en utilisant des tests statistiques dont on précisera l'hypothèse nulle et la statistique de test.

2 Régression hédonique

Pour établir un indice de prix immobilier, une méthode couramment employée est la méthode dite "hédonique". On considère que le prix d'un bien est la somme du prix de ses caractéristiques. Concrètement, cela revient à régresser les prix d'un bien sur ses variables explicatives. Pour obtenir un indice, nous allons ajouter des indicatrices temporelles et regarder l'évolution des coefficients.

2.1 Régression sans indicatrice temporelle

1. A partir des statistiques descriptives et de l'intuition économique, choisir un ensemble de variables explicatives. Expliquer le choix.
2. Estimer le modèle et commenter les résultats de la régression. Y a-t-il des surprises sur les valeurs des coefficients ? Précisez le cas échéant les statistiques de tests utilisées pour justifier votre réponse. Quel est le pouvoir explicatif du modèle ? Quelles en sont les limites ?

2.2 Création de l'indice

1. Justifier la méthode de création de l'indice.
2. En analysant le volume de données disponibles, choisir une fréquence temporelle appropriée.
3. Réaliser la régression, faire un graphe montrant l'évolution de l'indice, avec les intervalles de confiance correspondants. Commenter.

3 Prise en compte du biais de sélection

Pour obtenir un indice représentatif du parc immobilier, on souhaite corriger notre indice d'un éventuel biais de sélection, par exemple parce que les biens de moins bonne qualité mettent plus longtemps à se vendre. Pour cela, on va utiliser la procédure de Heckman, qui consiste dans un premier temps à évaluer la probabilité de vente d'un bien en utilisant un modèle probit puis à intégrer la probabilité estimée comme variable explicative dans notre régression (comme dans le TD 9/10).

3.1 Modèle de sélection

On a utilisé un modèle linéaire usuel :

$$Y = X'\beta + \epsilon$$

où X est le vecteur des caractéristiques pertinentes pour la détermination du prix. On suppose que notre sélection n'est pas aléatoire et qu'on peut modéliser S , la variable qui indique si le bien a été vendu, comme :

$$S = \mathbb{1}\{Z'\delta + \eta > 0\}$$

où Z est le vecteur des caractéristiques pertinentes pour la sélection, les caractéristiques du prix et éventuellement d'autres. Ainsi, on un bien est vendu si et seulement si $\eta > -Z'\delta$. Les hypothèses sont les suivantes :

1. (ϵ, η) est indépendant de Z et de moyenne nulle
2. $\eta \sim \mathcal{N}(0, 1)$
3. $\mathbb{E}(\epsilon|\eta) = \rho\eta$ avec $\rho \in \mathbb{R}$

La dernière équation signifie que les deux termes d'erreur sont corrélés linéairement. C'est cette hypothèse qui crée le biais de sélection : par exemple avec $\rho > 0$ les biens qui ont un η élevé (= plus de chances d'être vendus) auront ϵ élevé, donc on aura surtout des termes d'erreur élevés dans notre ensemble sélectionné.

1. Quelle pourrait être une raison du biais de sélection ? Expliquer la corrélation entre ϵ et η dans ce cas. Quel est le risque pour l'estimation de notre indice ?

Si on avait accès à η , on pourrait écrire :

$$\begin{aligned}\mathbb{E}(Y|Z, \eta) &= X'\beta + \mathbb{E}(\epsilon|Z, \eta) \\ &= X'\beta + \mathbb{E}(\epsilon|\eta) \\ &= X'\beta + \rho\eta\end{aligned}$$

Mais on connaît uniquement S ... Si on conditionne par S à la place, on obtient :

$$\begin{aligned}\mathbb{E}(y|Z, S = 1) &= \mathbb{E}[\mathbb{E}(Y|Z, \eta)|Z, S = 1] \\ &= \mathbb{E}[X'\beta + \rho\eta|Z, S = 1] \\ &= X'\beta + \rho\mathbb{E}(\eta|Z, S = 1) \\ &= X'\beta + \rho\mathbb{E}(\eta|Z, \eta > -Z'\delta)\end{aligned}$$

Mais comme on a fait l'hypothèse que $\eta \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}(\eta|Z, S = 1) = \frac{\phi(Z'\delta)}{\Phi(Z'\delta)} := \lambda(Z'\delta)$$

où $\lambda()$ est appelé l'inverse du ratio de Mills. Ainsi

$$\mathbb{E}(y|Z, S = 1) = X'\beta + \rho\lambda(Z'\delta)$$

On peut donc estimer le modèle en deux étapes : d'abord on trouve $\hat{\delta}$ puis on estime $\hat{\beta}$. Il y a un souci majeur avec cette méthode d'estimation : l'inverse du ratio de Mills est presque linéaire sur une large partie de son ensemble de définition. Donc si on a exactement les mêmes variables explicatives dans les deux étapes, i.e. $Z = X$, et que notre amplitude des X n'est pas suffisante pour sortir de la zone de linéarité de $\lambda()$, alors on aura une très forte multicollinéarité entre X et $\lambda(X'\delta)$ et de mauvaises erreurs standards.

2. Il nous faut trouver une variable instrumentale qui permette l'identification du biais de sélection pour que $Z \neq X$. Les principales raisons pour la mise en vente d'un bien sont les changements de structure familiale. Malheureusement, de telles caractéristiques ne figurent pas dans la base, mais nous disposons de l'âge des propriétaires. Expliquer en quoi l'âge des propriétaires peut servir de variable instrumentale.

Puisqu'on a une variable S dans $\{0, 1\}$ et qu'on a un terme d'erreur normal, η , il paraît naturel d'estimer δ avec un modèle de probit :

$$\mathbb{P}(S = 1|Z) = \Phi(Z'\delta)$$

3. Estimer un modèle probit pour le choix de mise en vente et commenter les résultats. L'âge est-il significatif ? Les valeurs des coefficients sont-elles cohérentes avec l'intuition ? Faire les tests statistiques correspondants en précisant à chaque fois l'hypothèse nulle et la statistique de test.

3.2 Correction de l'indice

1. Utiliser $\hat{\delta}$ pour calculer $\lambda(Z'\hat{\delta})$ pour chacune observation de la table de vente.

Même si on utilise $\hat{\delta}$ et pas δ , Heckman a montré que l'estimateur des MCO de β était convergent.

2. Recalculer l'indice en rajoutant la prédiction du ratio inverse de Mills dans les variables explicatives. Y a-t-il un biais de sélection significatif ? Faire les tests statistiques correspondants en précisant à chaque fois l'hypothèse nulle et la statistique de test.
3. Comme pour la question 2.2.3, réaliser un graphe de l'évolution de l'indice avec les intervalles de confiance correspondants.

4 Confrontation des résultats et conclusion

1. Mettre les deux indices sur un même graphe en les mettant tous les deux en base 100 pour la période initiale. Commenter les résultats.
2. Comparer avec l'indice des prix Île-de-France produit par l'Insee (disponible à l'adresse suivante : <https://www.insee.fr/fr/statistiques/serie/001587595>).
3. Conclure sur la méthodologie des indices de prix immobiliers (pertinence et limites).

5 (Question bonus)

Au lieu d'utiliser des indicatrices temporelles, on peut faire des régressions séparées pour chaque période et étudier l'évolution des estimations de prix d'un bien de référence. En choisissant le bien médian comme bien de référence, recalculer un indice de prix (pour les variables catégorielles, on prendra la modalité ayant un coefficient médian dans la régression globale). Commenter les résultats.

Fécondité et participation des femmes au marché du travail en Afrique sub-saharienne

Raphaël Lee
raphael.sh.lee@gmail.com

March 15, 2019

On s'intéresse à l'impact de la fécondité des femmes sur leur participation au marché du travail en Afrique sub-saharienne.

Données Les données proviennent des enquêtes démographiques et de santé et sont à télécharger sur le site du DHS : dhsprogram.com Il est nécessaire de s'inscrire pour pouvoir télécharger les données. Pour ce projet, c'est le questionnaire individuel femme qui sera utilisé (*Individual record*). Vous pouvez sélectionner le pays d'Afrique sub-saharienne de votre choix. Pour obtenir un nombre d'observations satisfaisant, vous empilerez toutes les vagues d'enquêtes disponibles.

Ressources Concernant la partie 4: Rivers, D., and Q. H. Vuong. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39: 347–366.

1. Statistiques Descriptives

1. Au vu des données : Quelle variable retiendrez-vous pour évaluer la participation au marché du travail ? Quelle variable pour la fécondité ?
2. Proposez des statistiques descriptives pertinentes et commentez-les.

2. Première modélisation

1. Proposez et estimez un modèle simple, sans variables de contrôles, de l'effet de la fécondité sur la participation au marché du travail. Commentez les résultats.
2. Quelles variables explicatives vous semble-t-il pertinent d'inclure dans le modèle et pourquoi ? Quelle variable est-il nécessaire d'inclure pour tenir compte du fait que l'échantillon de travail est constitué de 4 vagues empilées et pourquoi ?
3. Ecrivez le nouveau modèle comprenant des variables de contrôle supplémentaires et estimez-le. Commentez en comparant avec les estimations obtenues à la question 1. Effectuez les tests pertinents.

3. Problème d'endogénéité dans un modèle linéaire

1. L'hypothèse d'exogénéité de la variable de fécondité nécessaire à la validité du modèle utilisé dans la partie précédente vous semble-elle réaliste et pourquoi ?
2. Deux instruments communément utilisés dans la littérature sur la fécondité sont la naissance gémellaire de rang 1 (indicatrice qui vaut 1 si la femme a eu des jumeaux à la première naissance et 0 sinon) et le sexe des deux aînés (indicatrice qui vaut 1 si les deux aînés sont de même sexe et 0 sinon). Discutez de la validité de ces deux variables comme instruments.

Dans la suite on supposera la variable de fécondité continue.

3. Pour la suite, la naissance gémellaire de rang 1 sera utilisée comme instrument. Il faut alors restreindre l'échantillon aux femmes ayant au moins un enfant. Pourquoi ?
4. Ecrivez le modèle de la régression de première étape. Effectuez la régression de première étape et commentez. Quelle est la qualité de l'instrument ? Vous effectuerez les tests nécessaires pour en juger.
5. Estimez un modèle de probabilité linéaire par les doubles moindres carrés.

4. Problème d'endogénéité dans un modèle binaire

1. On souhaite maintenant estimer un probit avec variable instrumentale. La méthode à mettre en œuvre est celle développée par Rivers et Vuong (1989) et est présentée dans l'exercice 2 du TD 6-7. Ecrivez le modèle. Quelles sont les hypothèses nécessaires à la mise en œuvre de cette méthode ?
2. Effectuez l'estimation avec la méthode de Rivers et Vuong. Dans quelle mesure les résultats obtenus sont-ils comparables avec ceux obtenus avec un probit simple ? Expliquez.

5. Modélisation bi-probit

On considère maintenant une mesure binaire de la fécondité, à savoir l'indicatrice que le nombre d'enfant soit supérieur à s pour un certain seuil s . On considère le système suivant:

$$Y_1 = \mathbf{1}\{X'\beta_1 + Y_2\delta + \varepsilon \geq 0\}, \quad (1)$$

$$Y_2 = \mathbf{1}\{X'\beta_2 + Z\gamma + \eta \geq 0\}. \quad (2)$$

où Y_1 est l'indicatrice de participation au marché du travail, Y_2 correspond à l'indicatrice de fécondité définie précédemment et X sont les variables de contrôle considérés plus haut. (ε, η) est supposé suivre une loi normale bivariée $\mathcal{N}(0, \Sigma)$.

1. Discuter le choix de s et Z , et les restrictions que vous souhaitez imposer sur la matrice Σ . Ecrire la vraisemblance associée aux réalisations possibles de (Y_1, Y_2) . A quelle condition Y_2 est-elle endogène dans le système (1)-(2) ?
2. Estimer le modèle et comparer les résultats à ceux obtenus précédemment. Quel est l'intérêt de cette méthode par rapport aux précédentes ?