Problem Set #1

ECO 8000
Automne 2025

Date de remise: 27-oct-2025. Veuillez soumettre par courriel à `sam.gyetvay@gmail.com` avec ECO8000 + TP1 + votre nom de famille + numéro d'étudiant dans l'objet. Les remises en retard seront acceptées, mais je pourrais enlever des points.

Rappel: La collaboration est permise (et encouragée) mais chaque étudiant doit écrire et soumettre sa propre solution. Rappelez-vous que chaque devoir compte seulement pour 15% de votre note finale, tandis que la présentation + la participation + le projet final valent 70%. C'est volontaire: l'objectif est que vous appreniez, pas seulement que vous trouviez la bonne réponse. Si vous trouvez la bonne réponse en copiant d'un autre étudiant ou de ChatGPT, vous n'allez pas apprendre.

Instructions: Je voudrais que vous écriviez votre document avec LaTeX. Un compilateur LaTeX en ligne populaire parmi les étudiants et les chercheurs est Overleaf (`www.overleaf.com`). C'est une bonne pratique pour quand vous devrez écrire un article. Pour les questions qui impliquent l'analyse d'un base de données, vous pouvez utiliser le logiciel que vous voulez. J'utilise une combinaison de STATA et R dans mes recherches, et je suis assez avancé dans ma connaissance des deux, donc je peux vous donner des conseils. J'utilise parfois Python, mais je ne suis pas très bon. Je déconseille fortement d'utiliser MATLAB, SPSS, SAS, Julia ou tout autre logiciel. Pour les questions qui demandent d'écrire du code, veuillez inclure votre script (fichier STATA `.do`, script `.R`, ou autre). Ajoutez des commentaires dans votre code et indiquez quelle section correspond à chaque question pour que je puisse suivre.

Mot d'encouragement: Ce devoir est très long, et je vous ai donné beaucoup de temps pour le faire. Je vous encourage fortement à commencer *maintenant*. Travailler avec des données est probablement la compétence la plus précieuse que vous allez acquérir dans votre programme, et la seule façon de l'apprendre est d'y consacrer beaucoup de temps. Travailler avec des données est souvent pénible et frustrant. Vous passerez sûrement des nuits à crier devant votre ordinateur en essayant de déboguer du code et en jurant contre moi. On est tous passés par là. Respirez et essayez d'en profiter :-)

Question 1: Modèle de Roy généralisé (10 points) En cours, nous avons étudié une version du modèle de Roy où les travailleurs choisissent l'option 1 (par ex. chasser) plutôt que l'option 0 (par ex. pêcher) quand le revenu de 1 dépasse celui de 0: $Y_{i1} > Y_{i0}$. Dans ce problème, nous considérons une généralisation du modèle de Roy où le travailleur i choisit l'option 1 si $Y_{i1} - C_{i1} > Y_{i0} - C_{i0}$, où C_{i1} et C_{i0} sont des coûts.

Question 1 (a): (1 point) Nous commençons par définir

$$D_i = 1\{Y_{i1} - C_{i1} > Y_{i0} - C_{i0}\}$$

En mots: D_i est une variable indicatrice égale à 1 lorsque le travailleur i choisit l'option 1. Nous écrivons les revenus et les coûts comme fonctions d'un vecteur de caractéristiques observables du travailleur X_i:

$$Y_{id} = \beta_d X_i + \epsilon_{id}, \quad C_{id} = \gamma_d X_i + \eta_{id}, \quad d \in \{0, 1\}$$

où $E[\epsilon_{id}|X_i] = E[\eta_{id}|X_i] = 0$. Définissez $\psi = (\beta_1 - \beta_0) - (\gamma_1 - \gamma_0)$, et $v_i = (\eta_{i1} - \eta_{i0}) - (\epsilon_{i1} - \epsilon_{i0})$ et montrez que

$$D_i = 1\{\psi X_i > v_i\}$$

Décrivez cette condition en mots. Est-ce que les travailleurs choisissent toujours l'option 1 quand elle paie plus? Sinon, pourquoi pas?

Question 1 (b): (1 point) Montrez que

$$E[Y_{i1}|X_i, D_i = 1] = \beta_1 X_1 + E[\epsilon_{i1}|X_i, v_i < X_i]$$

Que représente $E[Y_{i1}|X_i, D_i = 1]$, en mots?

Question 1 (c): (2 points) Supposons que

$$(\epsilon_{id}, v_i)'|X_i \sim N\left(0, \begin{bmatrix} \sigma^2 & \rho_d\sigma_d \\ \rho_d\sigma_d & 1 \end{bmatrix}\right), \; d \in \{0, 1\}$$

Montrez que

$$E[Y_{i1}|X_i, D_i = 1] = \beta X_i + \rho_1\sigma_1\lambda_1(\psi X_i)$$

où $\lambda_1(z) = -\frac{\phi(z)}{\Phi(z)}$. Montrez aussi que

$$E[Y_{i0}|X_i, D_i = 0] = \beta_0 X_i + \rho_0\sigma_0\lambda_0(\psi X_i)$$

où $\lambda_0(c) = \phi(c)/[1 - \Phi(c)]$.

(*Indice*: utilisez l'astuce de la diapositive intitulée "Espérance conditionnelle de lois normales conjointes" pour écrire $\epsilon_{id} = \rho_d\sigma_d v_i + e_{id}$ où $E[e_{id}|v_i, X_i] = 0$)

Question 1 (d): (2 points) Supposons que nous avons une variable Z_i qui affecte les coûts de l'option 1 mais pas les revenus, de sorte que

$$Y_{i1} = \beta_1 X_i + \epsilon_{i1}, \quad C_{i1} = \gamma_1 X_i + \varphi_1 Z_i + \eta_{i1}$$

Nous estimons ensuite la relation entre D_i, X_i et Z_i en utilisant la méthode en deux étapes de

Heckman:

$$P(D_i = 1) = \Phi(\theta_1 X_i + \theta_2 Z_i) \tag{1}$$

$$Y_i = \theta_3 X_i + \theta_4 + \theta_5 \lambda_1(\hat{\psi}_1 X_i + \hat{\psi}_2 Z_i) \tag{2}$$

Quels sont les coefficients $\hat{\psi}_1$ et $\hat{\psi}_2$?

Question 1 (e): (2 points) Quels sont les paramètres structurels du modèle? Lesquels pouvons-nous identifier avec la méthode en deux étapes de Heckman dans la partie (d)?

Question 1 (f): (2 points) Supposons que D_i est la décision du travailleur i d'immigrer du pays d'origine vers le Canada, Y_{i1} les revenus au Canada, Y_{i0} les revenus dans le pays d'origine, et C_{i1} les coûts de migration. Quelle est la restriction d'exclusion à laquelle les variables X_i et Z_i doivent obéir? Expliquez-la en termes des équations, et aussi de façon intuitive en mots. Quelles sont quelques variables X_i et Z_i qui pourraient satisfaire la restriction d'exclusion?

Question 2: Mulligan et Rubinstein (2008) (20 points). Pour répondre à cette question, téléchargez le base de données au lien suivant

`www.samgyetvay.com/cps.html`

C'est le base de données analysé par Casey Mulligan et Yonah Rubinstein dans leur article de 2008 "Selection, Investment, and Women's Relative Wages Over Time."

Question 2 (a): (1 point) Nous commencerons par imposer les restrictions d'échantillon décrites dans l'article (section III.A). Restreignez l'échantillon aux travailleurs blancs, non hispaniques, âgés de 26 à 55 ans. Supprimez les travailleurs vivant en quartiers collectifs, et supprimez les travailleurs ayant une valeur manquante pour la variable `married`. (*Indice*: utilisez les commandes STATA `keep if` et/ou `drop if` pour retirer des observations de l'échantillon. Vous pouvez identifier les variables manquantes dans STATA en utilisant `missing()`.)

Question 2 (b): (1 point) Dans cette section, nous allons construire les variables utilisées dans la méthode en deux étapes de Heckman. Commencez par créer une variable indicatrice qui identifie les travailleurs "à temps plein toute l'année" : travailleurs ayant travaillé au moins 50 semaines et 35 heures par semaine. (*Indice*: utilisez les variables `wksly` et `hrslyr`). Créez une mesure du salaire horaire réel en divisant la variable `wage_cpi` par la variable `wrkhrlyr`, puis créez une variable égale au logarithme du salaire horaire réel. Ces deux variables seront les variables dépendantes dans le probit et la régression sur le salaire, respectivement. (*Indice*: utilisez la commande STATA `generate` pour créer de nouvelles variables.)

Question 2 (c): (3 points) Ensuite, nous allons générer les variables X et Z pour la méthode Heckman en deux étapes (section III.B). X contient des indicatrices pour différents niveaux d'éducation,

l'état civil, la région, et un polynôme d'ordre quatre ("quartique") en expérience professionnelle potentielle, interacté avec l'éducation. Commencez par créer une variable égale à 1 si le travailleur n'a jamais été marié, 0 sinon. Générez une variable indicatrice pour chacun des groupes d'éducation suivants : non diplômés du secondaire (8 ans de scolarité ou moins), non diplômés avec 9-11 ans de scolarité, diplômés du secondaire, travailleurs ayant un peu d'université, diplômés universitaires, diplômes avancés. Générez les composantes d'un polynôme d'ordre quatre en expérience (mesurée dans la variable `exp`) (*Indice*: pour éviter les problèmes de flottants avec de grandes valeurs, normalisez exp en la divisant par 10 avant de l'élever au carré, c.-à-d. créez `exp2 = (exp/10)^2`, `exp3 = (exp/10)^3`, etc.). Ensuite, créez un ensemble de variables nommées par exemple `exp1_hsd08`, `exp2_hsd08` pour chacune des interactions éducation-expérience. Enfin, utilisez la variable `state` pour définir une variable appelée `region` qui divise les états en nord-est, midwest, sud et ouest[1]. Z contient toutes les variables de X, plus une interaction entre l'état civil et le nombre d'enfants âgés de 0 à 6 ans. Le nombre d'enfants âgés de 0 à 6 ans est dans la variable `__age06`. Générez les termes d'interaction en créant des variables nommées par exemple `nmkids` égales à la variable indicatrice pour "jamais marié" multipliée par le nombre d'enfants âgés de 0 à 6 ans. Rassemblez toutes les variables de X et Z dans deux globals nommées `X` et `Z` (*Indice*: si vous n'êtes pas familier avec les globals, vous pouvez lire sur ce sujet ici `https://comet.arts.ubc.ca/docs/5_Research/econ490-stata/04_Locals_and_Globals.html`).

Question 2 (d): (2 points) Le nombre d'années d'expérience professionnelle n'est pas observé dans le CPS (Current Population Survey), sur lequel Mulligan et Rubinstein fondent leur analyse (et dont provient le fichier `cps.csv`). La variable `exp` est une mesure de l'expérience potentielle. Lisez l'Appendice I de l'article et expliquez avec vos mots comment les auteurs ont construit cette variable.

Question 2 (e): (3 points) Nous sommes maintenant prêts à estimer la première étape. Répétez le processus suivant, une fois pour la période 1975-79, et une fois pour la période 1995-99. Commencez par estimer par probit la relation entre "temps plein toute l'année", X et Z, pour les femmes seulement. Utilisez la variable `wgt` comme poids (option `[pw = wgt]`) (*Indices*: utilisez la commande STATA `probit`. Tapez `help probit` dans STATA pour plus d'informations. Vous pouvez utiliser les globals créés en (c) pour écrire chaque probit de manière compacte). Une fois le probit estimé, formez le terme du ratio de Mills inverse $\lambda(\hat{b}_1 X_i + \hat{b}_2 Z_i)$. Rappelez que $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$, où $\phi(z)$ est la densité normale standard et $\Phi(z)$ la densité cumulée standard. (*Indice*: dans STATA, utilisez `predict PhiXb, p` après `probit` pour générer $P(X_i \leq \hat{b}_1 X_i + \hat{b}_2 Z_i)$ et `predict Xb, xb` pour générer les valeurs ajustées $\hat{b}_1 X_i + \hat{b}_2 Z_i$. La commande `generate fXb = normd(Xb)` permet de calculer $\phi(\hat{b}_1 X_i + \hat{b}_2 Z_i)$). Fixez la valeur du ratio de Mills inverse à zéro pour les hommes.

Question 2 (f): (2 points) Avant d'estimer la seconde étape de Heckit, voyons ce que nous aurions si nous n'ajustions pas pour la sélection. Pour chaque période, réalisez une régression du loga-

[1] Voir `https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States` pour la liste des états de chaque région

rithme du salaire horaire réel sur une variable indicatrice égale à 1 si le travailleur est une femme, 0 sinon, et sur les covariables X, en restreignant l'échantillon aux travailleurs employés à temps plein toute l'année 50 semaines par an. Utilisez des erreurs standards robustes à l'hétéroscédasticité. Reportez le coefficient de la variable indicatrice femme. Comment interprétez-vous ce nombre? Comment évolue-t-il entre les deux périodes? Quelles implications pour l'écart salarial entre hommes et femmes? Pourquoi devrions-nous être sceptiques à propos de cette estimation? (*Indice*: utilisez la commande `regress` pour les régressions linéaires et l'option `, robust` pour des erreurs standards robustes.)

Question 2 (g): (4 points) Nous sommes maintenant prêts à estimer la seconde étape de Heckit. Estimez la même régression que dans (f), mais incluez cette fois la variable du ratio de Mills inverse comme contrôle. Reportez les coefficients sur la variable indicatrice femme ainsi que le coefficient sur le ratio de Mills inverse. Interprétez chaque coefficient. Comment les coefficients de la variable femme se comparent-ils à ceux de la partie (f)? Les femmes sont-elles sélectionnées négativement ou positivement? Pourquoi? Comment cela a-t-il changé entre 1975-79 et 1995-99?

Question 2 (h): (4 points) Répétez les parties (f) et (g) séparément pour les femmes actuellement mariées et celles qui ne le sont pas. Regroupez vos observations dans un tableau et commentez les résultats.

Question 3: Angrist et Krueger (1991) (20 points). Pour répondre à cette question, téléchargez le fichier `NEW7080.rar` sous **Angrist et Krueger (1991)** au lien suivant

`https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive`

(Ceci est un fichier `.rar`, ce qui signifie que vous devez en extraire le contenu). C'est le base de données analysé par Josh Angrist et Allan Krueger dans leur article de 1991 "Does compulsory school attendance affect schooling and earnings?"

Question 3 (a). (2 points) Ce base de données n'est pas très propre. Les noms des variables sont simplement `v1`, `v2`, etc. Les fichiers STATA `.do` `QOB_Table_IV.do`, `QOB_Table_V.do`, et `QOB_Table_VI.do` au lien ci-dessus contiennent du code STATA qui permettra d'identifier les noms des variables. Utilisez-les pour nettoyer les données. Essayez de deviner ce que chaque variable représente en vous basant sur son nom. Il peut être utile de jeter un coup d'oeil à l'article.

Question 3 (b). (1 point) La variable `YOB` contient l'année de naissance de chaque individu du base de données. Cependant, cette variable n'est pas codée de façon cohérente selon les années. Transformez la variable pour qu'elle soit cohérente. (*Indice*: utilisez la commande STATA `tabulate YOB` pour résumer les valeurs).

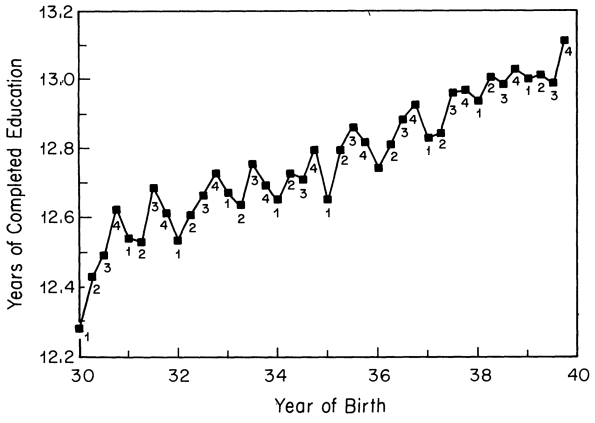

FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

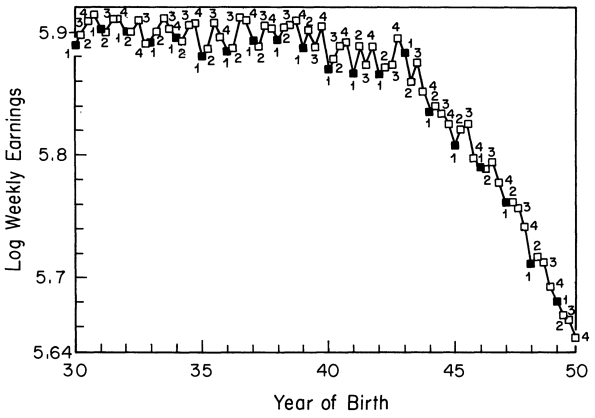

FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Question 3 (c). (5 points) Recréez les Figures I, II et V d'Angrist et Krueger (1991). Essayez de rendre la figure aussi proche que possible de celle de l'article. Expliquez comment le motif des figures se rapporte à la stratégie d'identification de l'article. (*Indice*: vous pouvez combiner les deux variables `YOB` et `QOB` en créant une nouvelle variable égale à `YOB + QOB/4`). (*Indice*: utilisez les commandes STATA `twoway line` et `twoway scatter`).

Question 3 (d). (3 points) Le base de données contient les recensements de 1970 et 1980. Utilisez ces observations pour créer une version des Figures I-II et V pour les travailleurs nés entre 1920 et 1929. (*Indice*: la variable `LWKLYWGE` a des valeurs manquantes pour tous les travailleurs nés entre 1920 et 1929. Cependant, l'une des variables non nommées (c.-à-d. l'une de `v14`, `v15`, `v17`, etc.) contient des valeurs non manquantes.)

Question 3 (e). (5 points) Reproduisez Table III de l'article. Ne vous inquiétez pas de le faire exactement identique à celui de l'article. (*Indice*: utilisez la commande `summarize` pour calculer les moyennes. Pour les différences dans la colonne (3), utilisez la commande `regress` et régressiez le résultat sur une variable indicatrice égale à 1 si le travailleur est né au premier trimestre). Vous n'avez pas besoin de calculer les erreurs standards pour l'estimation de Wald, mais je donnerai des points bonus si vous pouvez les calculer et si elles sont identiques à celles de l'article.

Question 3 (f). (4 points) Répétez l'analyse du Panel B du Table III pour les individus nés entre 1940-49. Commentez vos résultats.

Question 4 : Théorème LATE. (20 points) Dans ce problème, vous allez explorer le théorème LATE, prouvé pour la première fois par Joshua Angrist et Guido Imbens dans leur article de 1994 "Identification and estimation of local average treatment effects" à travers un exemple hypothétique. Une organisation à but non lucratif dans le domaine de l'éducation a collaboré avec des économistes du travail à l'UQAM pour tester un programme visant à sensibiliser les étudiants aux

subventions de frais de scolarité pour les familles à faible revenu. Ensemble, l'équipe de recherche a constitué un échantillon de 1000 jeunes adultes âgés de 18 à 21 ans dont le revenu parental est inférieur à 150% du seuil de pauvreté. Les chercheurs ont assigné aléatoirement chaque étudiant i au groupe de traitement ($Z_i = 1$) ou au groupe de contrôle ($Z_i = 0$). Les étudiants du groupe de traitement reçoivent un dossier d'information expliquant les subventions auxquelles ils ont droit et des instructions détaillées sur la façon de postuler. Soit D_i une variable indicatrice égale à 1 si i s'inscrit à un programme et 0 sinon. Et soit Y_i le revenu de i cinq ans plus tard.

Question 4 (a). (5 points) Soit $D_i(1)$ l'inscription du travailleur i s'il reçoit l'information, et $D_i(0)$ sa décision d'inscription s'il n'a pas reçu l'information. Décrivez en mots les groupes suivants, dans le contexte de cette application :

- $1 = D_i(1) > D_i(0) = 0$
- $D_i(1) = D_i(0) = 1$
- $D_i(1) = D_i(0) = 0$

Question 4 (b). (5 points) Énoncez formellement les trois hypothèses du théorème LATE (issues des slides) et expliquez-les avec vos mots dans le contexte de cette application. Pour chaque hypothèse, expliquez pourquoi elle est satisfaite (ou non) dans ce cas.

Question 4 (c). (5 points) Prouvez le théorème LATE. C'est-à-dire, montrez que, sous les hypothèses énoncées à la partie (b),

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$$

Question 4 (d). (5 points) Attendez-vous à ce que le LATE soit plus grand, plus petit ou égal à l'ATE dans cette application ? Expliquez votre réponse.

Question 5 : Acemoglu, Autor et Lyle (2004). (20 points) Dans cette question, vous allez reproduire certaines des figures et tableaux principaux de l'article "Women, War and Wages: The Effect of Female Labor Supply on the Wage Structure at Mid-Century" de David Autor, Daron Acemoglu et David Lyle. Pour télécharger les données, rendez-vous sur le site suivant :

`https://economics.mit.edu/people/faculty/daron-acemoglu/data-archive`

faites défiler jusqu'en bas pour trouver le titre de l'article, cliquez dessus et téléchargez certains bases de données. Contrairement aux autres problèmes appliqués, pour lesquels je vous ai donné des instructions étape par étape et des indications détaillées, vous êtes cette fois-ci autonome.

Sur le lien ci-dessus, vous trouverez des fichiers STATA `.do` qui vous seront utiles pour votre reproduction. Il peut également être utile de lire l'article, en particulier les sections I et III.A.

Question 5 (a). (5 points) Reproduisez les Figures 3, 4, 5, 6, 7, 8 et 9. Faites-les ressembler le plus possible aux figures de l'article, y compris les initiales des états comme étiquettes. Pour chaque figure, discutez avec vos mots de la manière dont elle se rapporte à l'argument de l'article.

Question 5 (b). (5 points) Reproduisez le Tableau 3. Pourquoi les estimations OLS ne donnent-elles pas d'information sur la forme de la courbe de demande de travail ?

Question 5 (c). (5 points) Reproduisez les Tableaux 5 et 6. Résumez brièvement. Pourquoi est-il important de contrôler la fraction d'hommes dans les professions et les industries en 1940 ?

Question 5 (d). (5 points) Reproduisez les Tableaux 7, 8 et 9. Résumez-les brièvement.