TSE
Toulouse School of Economics
École d'économie de Toulouse

UNIVERSITÉ TOULOUSE 1 CAPITOLE

**Université Toulouse 1 Capitole**
**Ecole d'économie de Toulouse**

**Année universitaire 2016-2017**

**Session 1**

**Semestre 2**

| |
|---|
| Master 1 Economics, Econometrics & Statistics, Economie & droit |
| Epreuve : Program Evaluation |
| Date de l'épreuve : 27 mars 2017 |
| Durée de l'épreuve : 2h |

Liste des documents autorisés : Aucun

Liste des matériels autorisés : Calculette

Nombre de pages (y compris page de garde) : 7

**L'examen se compose de trois parties indépendantes. Vous pouvez répondre dans l'ordre que vous préférez.**

**PARTIE A: 6 points.** Répondez aux questions suivantes. Définissez les variables si vous utilisez des formules mathématiques.

**Question 1: (2 pts) One-sided compliance et LATE**

a. (0.5 pts) Expliquer ce qu'est le problème de one-sided compliance?

b. (0.25 pts) Que sont les compliers et non-compliers?

c. (0.5 pts) Définir le local average treatment effect (LATE) avec des mots et donnez sa formule mathématique.

d. (0.75 pts) Montrer que le local average treatment effect (LATE) est égal à l'intention to treat pour les compliers ($\text{ITT}_{CO}$).

**Question 2: (2 pts) Design de regression discontinuity**

a. (0.5 pts) Expliquer la différence entre un sharp et un fuzzy design de regression discontinuity.

b. (0.75 pts) Quel effet de traitement peut-on identifier avec une approche par regression discontinuity? Donner l'expression de cet effet de traitement pour un sharp design de regression discontinuity et pour un fuzzy design de regression discontinuity.

c. (0.75 pts) On veut estimer l'effet du traitement D (0 ou 1) sur l'outcome Y. Le statut de traitement est 1 quand la forcing variable X supérieure ou égale au seuil c. En s'appuyant sur les Figures 1-3, expliquez pourquoi une approche de regression discontinuity serait appropriée (justifiez et précisez quelle figure apporte quelle information).

**Question 3: (2 pts) Two-sided compliance et LATE**

a. (0.25 pts) Donner un example d'experience où le problème de two-sided compliance pourrait se poser.

b. (0.75 pts) Quelles hypothèses sont nécessaires pour identifier le LATE quand on a un problème de two-sided compliance?

c. (0.75 pts) Définir les concepts de validité interne et validité externe. Que peut-on dire de la validité interne et de la validité externe du LATE?

d. (0.25 pts) On veut estimer l'effet d'un traitement D (0 ou 1) sur Y. Soit $Z \in \{0, 1\}$ l'assignement du traitement. En utilisant les résultats des deux régressions suivantes, quel est le local average treatment effect (LATE)?

$$Y = 0.1 + 3 \times Z + \varepsilon_y$$
$$D = 0.2 + 0.3 \times Z + \varepsilon_d$$

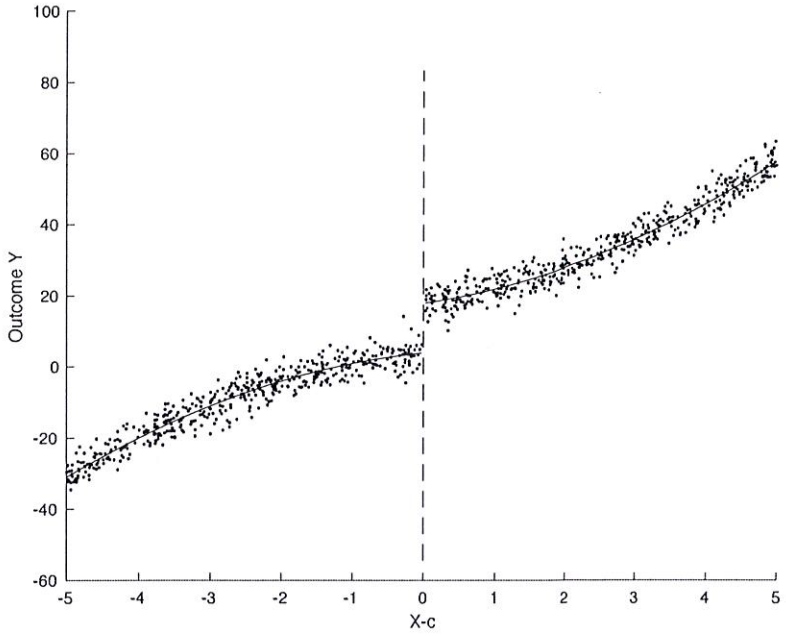


Figure 1: Évolution de l'outcome (Y) en fonction de la forcing variable (X) moins le seuil (c)

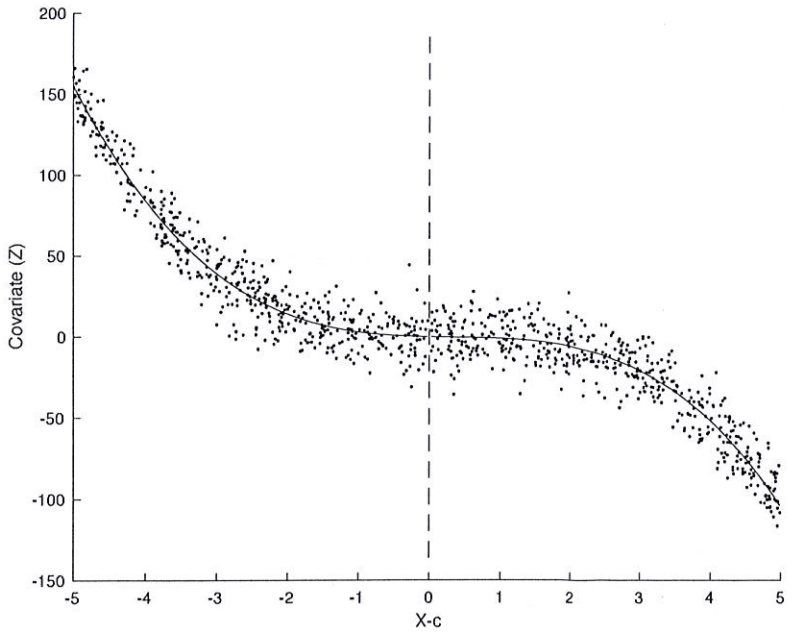


Figure 2: Évolution d'une variable extérieure $Z$ en fonction de la forcing variable (X) moins le seuil (c)

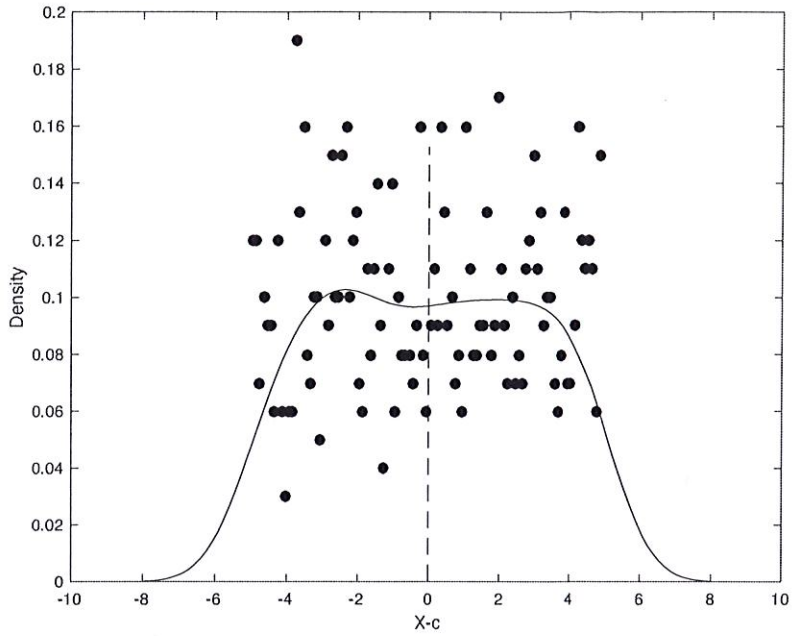


Figure 3: Densité de la forcing variable (X) moins le seuil (c)

**PARTIE B: 8 points.**

1. **(0.75 pt)** Expliquez clairement le paradoxe de Simpson.

2. **(1 pt)** On note la fonction de propensity score par $\pi(x)$, avec $\pi(x) \in (0,1)$. Par simplicité, nous supposons que la variable $X$ is univariée. Montrez que

$$E(X \mid D = 0, \pi(X) = \pi) = E(X \mid D = 1, \pi(X) = \pi).$$

3. **(2 pts)** Dans la suite, nous avons de la sélection sur observables et nous considérons des méthodes non-paramétriques basées sur les noyaux (kernels). Supposons que l'on dispose d'un échantillon d'observations, $i = 1, ..., n$.

   (a) (0.5 pt) Donnez un estimateur non-paramétrique de $E(D \mid X = x)$.

   (b) (0.5 pt) Donnez un estimateur non-paramétrique de $E(Y \mid X = x, D = d)$ avec $d = 0, 1$.

   (c) (0.5 pt) Donnez un estimateur non-paramétrique de $ATT(x)$.

   (d) (0.5 pt) Quel est l'impact sur cet estimateur du choix d'une fenêtre (bandwidth) trop petite? D'une fenêtre trop grande? Comment procéder en pratique pour choisir cette fenêtre?

4. **(1 pt)** Nous avons encore de la sélection sur observables. Nous savons que sous certaines conditions nous avons

$$ATT = \frac{1}{\Pr(D = 1)} \, E\left( \frac{D - \pi(X)}{1 - \pi(X)} Y \right).$$

(a) (0.5 pt) Donnez un estimateur non-paramétrique de ATT.

(b) (0.5 pt) Quel est le défaut de cet estimateur?

5. **(3.25 pts)** Considérons une variable de traitement $D$ qui dépend uniquement de la variable $X$. Supposons que l'outcome potentiel est donné par

$$Y(d) = \alpha(d) + \beta(d)X + \gamma Z + U(d), \quad E(U(d) \mid X, Z, D = d) = 0,$$

où la variable $Z$ est supposée indépendante de $X$ et $D$, alors que $\alpha(0), \alpha(1), \beta(0), \beta(1), \gamma$ sont non-aléatoires avec $\gamma \neq 0$.

(a) (0.5 pt) Considérez la régression:

$$Y_i = a + X_i\, b + Z_i\, c + D_i\, \tau + D_i X_i\, e + Z_i D_i\, f + \varepsilon_i.$$

Caractérizez les paramètres $a, b, c, \tau, e, f$ en fonction $\alpha(d), \beta(d), d = 0, 1$ et $\gamma$.

(b) (0.5 pt) Quel est le paramètre ATE? Comment l'estimer à partir de la régression précédente?

(c) (0.75 pt) Considérez maintenant la régression

$$Y_i = \tilde{a} + X_i\, \tilde{b} + D_i\, \tilde{\tau} + D_i X_i\, \tilde{e} + \tilde{\varepsilon}_i.$$

Ceci est connu comme un problème d'oubli de variable. Caractérisez les paramètres $\tilde{a}, \tilde{b}, \tilde{\tau}, \tilde{e}$ en fonction de $a, b, c, \tau, e, f$.

(d) (0.5 pt) Quel est le paramètre ATE? Comment l'estimer à partir de la régression précédente?

(e) (1 pt) Parmi ces deux estimateurs de ATE, quel est le meilleur? Pourquoi?

**PARTIE C: 6 points.**

Nous voulons savoir si les employés du secteur public ont un salaire plus élevé que les employés du secteur privé.

Nous utilisons un ensemble de données extrait de l'Enquête emploi en France en 2011 sur le secteur public et le secteur privé (définissant la variable $Pub = 1$ si la personne est employée par le secteur public, 0 sinon). Nous observons aussi le taux de salaire (en logarithmes, $y$), le niveau d'éducation (en trois groupes: collège, lycée et supérieur), l'âge en trois groupes (25-35, 36-45, 46-55) et le sexe (s = 1 pour les hommes).

Nous construisons des cellules en croisant les niveaux d'éducation, les groupes d'âge et la variable homme/femme. Cette construction fournit 3x3x2 = 18 cellules. Dans chaque cellule, on estime le nombre d'observations ($N_c$), la probabilité de travailler dans le secteur public ($\hat{p}_c$) et la différence du salaire moyen (en log) entre le secteur public et le secteur privé , $\bar{y}_{c1} - \bar{y}_{c0}$ en utilisant par exemple:

$$\bar{y}_{c1} = \frac{1}{N_{c1}} \sum_{i \in c, Pub_i = 1} y_i,$$

où $N_{c1} = N_c \hat{p}_c$ est le nombre d'employés publics dans la cellule $c$. Les résultats sont donnés dans le tableau 1.

On régresse également le logarithme du taux de salaire sur l'indicateur du secteur public et dans cette régression brute, on obtient pour le coefficient de l'indicateur du secteur public, une estimation de 0,047 avec un ecart-type (s.e.) égal à 0,0085.

**Partie I: 3 pts**

**Question 1:** (0.5 pts) Définir précisément ce qu'est le traitement, quelles sont les outcomes potentiels et quel est l'effet moyen du traitement sur les traités (ATT). Expliquer avec des mots ou par des formules.

**Question 2:** (0.5 pts) Sous quelle condition, la dernière colonne, $\bar{y}_{c1} - \bar{y}_{c0}$, fournit une estimation de l'ATT dans chaque cellule d'éducation, âge et sexe?

**Question 3:** (1 pt) L'hypothèse de support commun est-elle satisfaite? Argumentez en particulier en utilisant des arguments statistiques.

**Question 4:** (0.5 pts) Décrire la régression paramétrique qui conduirait à des estimations de l'ATT dans toute la population lorsque les outcome potentiels sont donnés par:

$$Y(d) = X\beta(d) + \varepsilon(d)$$

où X comprend une constante, les indicateurs de deux (sur trois) variables d'éducation, deux (sur trois) groupes d'âge et une variable de sexe.

**Question 5:** (0.5 pts) Décrire l'estimation de l'ATT en fonction des estimations de $\hat{\beta}(d)$. Par cette méthode, on obtient $\hat{\tau}_{Reg}$ =.034 (s.e. 0077). Comparer avec la régression brute en termes de biais de sélection. En utilisant la dernière colonne du tableau, quelles sont les variables responsables du biais?

**Partie II: 3 pts**

**Question 6:** (0.5 pts) Pour une observation traitée dans une cellule $c$ définie ci-dessus, quel est le groupe naturel de contrôle?

**Question 7:** (0.5 pts) Expliquez pourquoi l'appariement ("matching") multiple doit être privilégié dans ce cas.

**Question 8:** (0.5 pts) Expliquez l'arbitrage derrière le choix d'apparier avec remplacement ou sans remplacement.

**Question 9:** (1 pt) Dériver l'expression de l'ATT en utilisant cette procédure d'appariement en fonction de $N_c$, $\hat{p}_c$ et $\bar{y}_{c1} - \bar{y}_{c0}$ définis ci-dessus. On obtient $\hat{\tau}_{Matching} = .033$ (s.e. 0.0077)

**Question 10:** (0.5 pts) Pourquoi y a-t-il une différence entre $\hat{\tau}_{Matching}$ et $\hat{\tau}_{Reg}$? Pourquoi est-elle si petite?

Table 1: Charactéristiques par cellule

| Educ | Age | Sexe | Nombre | Public | Diff salaires |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 49 | 0.16 | 0.11 |
| | | | | (0.053) | (0.073) |
| 1 | 1 | 1 | 96 | 0.062 | -0.13 |
| | | | | (0.025) | (0.09) |
| 1 | 2 | 0 | 137 | 0.18 | -0.019 |
| | | | | (0.033) | (0.047) |
| 1 | 2 | 1 | 175 | 0.15 | -0.054 |
| | | | | (0.027) | (0.042) |
| 1 | 3 | 0 | 294 | 0.28 | 0.0076 |
| | | | | (0.026) | (0.026) |
| 1 | 3 | 1 | 268 | 0.097 | -0.023 |
| | | | | (0.018) | (0.052) |
| 2 | 1 | 0 | 444 | 0.24 | 0.058 |
| | | | | (0.02) | (0.018) |
| 2 | 1 | 1 | 608 | 0.11 | -0.016 |
| | | | | (0.013) | (0.028) |
| 2 | 2 | 0 | 623 | 0.28 | 0.0065 |
| | | | | (0.018) | (0.019) |
| 2 | 2 | 1 | 677 | 0.16 | -0.0082 |
| | | | | (0.014) | (0.027) |
| 2 | 3 | 0 | 744 | 0.29 | 0.048 |
| | | | | (0.017) | (0.02) |
| 2 | 3 | 1 | 733 | 0.17 | -0.015 |
| | | | | (0.014) | (0.026) |
| 3 | 1 | 0 | 733 | 0.28 | 0.092 |
| | | | | (0.017) | (0.023) |
| 3 | 1 | 1 | 492 | 0.18 | -0.037 |
| | | | | (0.017) | (0.035) |
| 3 | 2 | 0 | 581 | 0.38 | 0.066 |
| | | | | (0.02) | (0.028) |
| 3 | 2 | 1 | 376 | 0.21 | 0.059 |
| | | | | (0.021) | (0.038) |
| 3 | 3 | 0 | 285 | 0.39 | 0.092 |
| | | | | (0.029) | (0.038) |
| 3 | 3 | 1 | 191 | 0.31 | 0.017 |
| | | | | (0.033) | (0.058) |

Note: "Educ", "Age" et "Sexe" sont les valeurs des covariables définies ci-dessus.

"Nombre"$=N_c$, le nombre d'observations par cellule; "Public" est $\hat{p}_c$, la proportion d'employés du public; "Diff salaires" est $\bar{y}_{c1} - \bar{y}_{c0}$, la différence entre la moyenne des logarithmes des salaires dans le secteur public et le secteur privé dans chaque cellule. Ecart-types des estimations entre parenthèses.