Devoir Maison

Pratiques de la Recherche en Économie, CPES3

3 Décembre 2025

Consignes

- Le DM est à rendre le 8 Janvier 2026 avant 23h59
- Le rendu consistera en un fichier pdf qui contiendra vos réponses (au format nom_prenom.pdf) et un script `R` (au format nom_prenom.R ou nom_prenom.Rmd)
- Vous pouvez le faire seul/seule ou en binôme
- Veillez à commenter votre code
- La base de données à utiliser est `simulated_sibling_data.rds`, téléchargeable sur le Moodle
- Veillez à respecter à la lettre le nom des dataframes et variables suggéré. **1 point sera retiré au DM si cette consigne n'est pas respectée.**

Données

Les données ont été construites à partir de distributions empiriques de différentes variables issues de l'enquête *Formation et Qualification Professionnelle (FQP)* de l'INSEE, qui n'est malheureusement pas disponible en open data. Elles ont été **entièrement simulées** de manière à se rapprocher des distributions marginales et des dépendances entre variables observées dans l'enquête. **Cependant, la simulation reste approximative et les ordres de grandeur ne correspondent pas exactement aux valeurs observées dans l'enquête réelle.**

Le dictionnaire des variables se trouve page 5.

Objectif de ce DM

L'objectif de ce DM est de quantifier l'importance de la taille de la fratrie sur la réussite scolaire.

Partie 1: Nettoyage de données (7pts)

Cette première partie vise à nettoyer et préparer les données en vue des analyses descriptives et économétriques des parties suivantes.

Importez sur `R` la base de données que vous nommerez `df`.

Question 1 (5pts):

Créez les variables suivantes:
- `cohorte` qui regroupe les années de naissance en cohortes de 5 ans (par exemple : "1945-1949", etc).
- `femme`, une dummy égale à 1 si l'individu est une femme, 0 sinon.
- `rang_naissance` qui indique le rang de naissance de chaque individu au sein de sa fratrie (1 pour l'aîné, 2 pour le deuxième enfant, etc.).
- `elder`, une dummy égale à 1 si l'individu est l'aîné de sa fratrie, 0 sinon.
- `cohorte_aine` qui indique la cohorte de naissance de l'aîné pour chaque fratrie.
- `taille_fratrie` qui indique le nombre d'enfants au sein de chaque fratrie.
- `plus_de_2`, une dummy égale à 1 si la fratrie comprend 3 enfants ou plus, 0 sinon.
- `compo_genre` qui décrit la composition par sexe des deux premiers enfants de chaque fratrie. Cette variable prend quatre valeurs possibles: "FF" (deux filles), "MM" (deux garçons), "FM" (fille puis garçon), ou "MF" (garçon puis fille).

Question 2 (0,5pt):

Recodez les variables `mother_cs` et `father_cs` en remplaçant les codes numériques (1 à 6) par les labels des catégories socioprofessionnelles correspondantes.

Question 3 (1,5pt):

- Représentez la distribution de la taille de la fratrie. Commentez.
- Supprimer les fratries dont la taille fait parti des 1% plus élevées.

Partie 2: Statistiques Descriptives (8pts)

Cette partie a pour objectif d'explorer les données et comprendre les déterminants du niveau de diplôme.

Partie A: Taille de la fratrie

Question 4 (0,5pt):

Représentez graphiquement l'évolution de la taille de la fratrie selon la cohorte de naissance de l'aîné. Commentez.

Question 5 (1pt):

Représentez graphiquement l'évolution de la part de fratries de 3 enfants ou plus selon la cohorte de naissance de l'aîné en distinguant les quatre configurations possibles du genre des deux aînés. Commentez.

Question 6 (1,5pt):

Représentez l'évolution de la taille moyenne des fratries selon la PCS du père. Selon vous, quelles raisons peuvent expliquer la différence de fécondité entre les différents groupes sociaux et son évolution?

Partie B: Origine sociale

Question 7 (2pts):

Représentez l'évolution du niveau de diplôme moyen selon la CSP de la mère, en ne conservant que les individus ayant un niveau de diplôme connu. Commentez.

Partie C: Caractéristiques individuelles

Question 8 (2pts):

Représentez graphiquement le niveau de diplôme moyen selon le rang de naissance, en ne conservant que les rangs jusqu'à 5 maximum et les individus ayant un niveau de diplôme connu. Commentez. Selon vous, quels mécanismes peuvent expliquer cela?

Question 9 (1pt):

Représentez graphiquement l'évolution du niveau de diplôme moyen par cohorte de naissance selon le sexe. Commentez.

Partie 3: Analyse Économétrique (11pts)

Cette analyse économétrique explore les déterminants du niveau de diplôme.

Question 10 (1pt):

Selon vous, la taille de la fratrie a t-elle un effet positif, négatif, ou nul sur le niveau de diplôme? Pourquoi?

Question 11 (1,5pt):

- Régressez le niveau de diplôme sur la taille de fratrie. Vous nommerez ce modèle `reg1`.
- Régressez le niveau de diplôme sur la dummy `plus_de_2`. Vous nommerez ce modèle `reg2`.
- Dans les deux cas, que représentent $\hat{\alpha}$ et $\hat{\beta}$? Peut-on dire que $\hat{\beta}$ représente l'effet causal de la taille de la fratrie sur le niveau de diplôme? Pourquoi?

Question 12 (1,5pt):

Ajoutez le fait d'être une femme, l'aîné ainsi que la cohorte de naissance dans le modèle précédemment estimé (`reg1`), que vous nommerez `reg3`. Commentez.

Question 13 (1pt):

Ajoutez la catégorie socio-professionnelle du père et de la mère dans le modèle précédent que vous nommerez `reg4`. Que constatez-vous?

Question 14 (1pt):

Régressez le niveau de diplôme sur l'interaction entre `femme` et `elder` ainsi que la taille de la fratrie et la CSP des parents. Nommez ce modèle `reg5`. Commentez.

Question 15 (1pt):

Reprenez le modèle de la question 13 en remplaçant `elder` par le rang de naissance et nommez-le `reg6`. Commentez.

Question 16 (2pts):

- Régressez la taille de la fratrie sur la variable `compo_genre`, le fait d'être une femme, l'aîné, la cohorte de naissance et la PCS des parents. Nommez ce modèle `reg7`. Commentez uniquement les effets observés de la variable `compo_genre`.
- En utilisant la fonction `ivreg` du package du même nom, estimez l'effet de la taille de la fratrie en utilisant comme instrument `compo_genre` et nommez ce modèle `reg8`. Commentez.

Question 17 (2pts):

Selon-vous, le modèle précédent permet-il d'estimer l'effet causal de la taille de la fratrie sur le niveau de diplôme?

Dictionnaire des variables

Nom	Variable	Description
Identifiant famille	`family_id`	Identifiant unique de la fratrie
Année de naissance	`birth_year`	Année de naissance de l'individu
Sexe	`sexe`	Sexe de l'individu 1 = Garçon 2 = Fille
CSP du père	`father_cs`	Catégorie socioprofessionnelle du père 1: Agriculteur 2: Artisan 3: Cadre 4: Profession intermédiaire 5: Employé 6: Ouvrier
CSP de la mère	`mother_cs`	Catégorie socioprofessionnelle de la mère 1 à 6, mêmes modalités
Diplôme	`dipl`	Niveau de diplôme 1: Aucun diplôme 2: Diplôme < Baccalauréat 3: Baccalauréat 4: Bac+2 5: > Bac+2