

Econométrie appliquée sur Stata

APPLICATIONS

Septembre 2026

Application 1.

Ajout d'observations

L'objectif de cet exercice est d'apprendre à empiler deux bases de données.

Vous avez à votre disposition la base de données `individus07_silc` correspondant à l'information au niveau des individus à partir des données françaises de l'enquête européenne EU-SILC sur le revenu et les conditions de vie.

Après avoir ouvert et enregistré un dofile, renseigné le chemin de votre répertoire de travail et ouvert la base de données, on vous demande de :

Q.1. Tabuler la variable associée au sexe des individus

Q.2. Identifier le code associé à chaque sexe

Q.3. Créer 2 bases de données distinctes :

- Une base contenant uniquement les informations reportées par des femmes :
`base_travail_femmes.dta`
- Une base contenant uniquement les informations reportées par des hommes :
`base_travail_hommes.dta`

Q.4. Empiler les fichiers `base_travail_femmes.dta` et `base_travail_hommes.dta`

Application 2.

Fusion de bases de données

L'objectif de cet exercice est de fusionner deux bases (au niveau individuel et au niveau ménage) afin d'obtenir une base complète rassemblant toutes les informations issues de ces dernières. On souhaite apparier l'information au niveau ménage à l'information au niveau des individus à partir des données françaises de l'enquête européenne EU-SILC sur le revenu et les conditions de vie.

Vous avez à votre disposition deux bases :

- la base individu : `individus07_silc`
- la base ménage : `menages07_silc`

Après avoir ouvert et enregistré un dofile, puis renseigné le chemin de votre répertoire de travail, on vous demande de procéder aux opérations suivantes :

Q.1. Ouvrir la base individus "`individus07_silc.dta`" et identifier le nombre d'individus et de ménages

Q.2. Fusionner la base individus avec la base ménages "`menages07_silc.dta`" et justifier la commande utilisée

Q.3. Vérifier et commenter le résultat de la fusion

Q.4. Enregistrer la base sous le nom "`individus_menages07_silc_travail.dta`"

Q.5. Référez-vous au dictionnaire des codes pour trouver la variable représentant le salaire net. Utilisez cette variable pour calculer le salaire net total par ménage.

Q.6. Créez une variable indiquant le nombre d'individus par ménage. En moyenne, combien de membres d'un ménage sont interrogés ?

Application 3.

Régressions linéaires et tests d'hypothèses

Dans cet exercice, on s'intéresse au lien entre éducation et revenus issus du travail. Pour cela, on utilisera des données en coupe transversale issues de l'enquête nationale réalisée auprès des ménages brésiliens en 2015 (PNAD). Deux fichiers de données sont mis à disposition :

- Des données au niveau individuel : `PNAD2015_individuals.dta`
- Des données au niveau ménage : `PNAD2015_households.dta`

Le questionnaire ainsi que le dictionnaire des codes vous sont également fournis.

Après avoir ouvert et enregistré un dofile, puis renseigné le chemin de votre répertoire de travail, on vous demande de procéder aux opérations suivantes :

Q.1. Ajouter l'information au niveau des ménages aux données individuelles. On enregistrera au préalable les base de travail sous les noms « `work_pnad2015_ind.dta` » et « `work_pnad2015_hh.dta` ». Conserver uniquement les lignes pour lesquelles la fusion a bien été réalisée. Le résultat de la fusion sera lui enregistré sous « `work_pnad2015_merged.dta` » et sera utilisé dans la suite de l'exercice.

Q.2. Effectuer les opérations suivantes pour préparer les variables qui nous serviront ensuite :

- a) Garder uniquement les 15 ans et plus dans la base de travail.
- b) Créer la variable `woman` qui prendra la valeur 1 pour les femmes, et 0 pour les hommes.
- c) Créer les variables `age` et `age_sq`, correspondant respectivement à l'âge et à l'âge au carré de l'individu.
- d) Créer une variable dichotomique nommée `work` égale à 1 si l'individu a travaillé au cours de la dernière semaine, 0 sinon.
- e) Créer une variable nommée `wage` correspondant au salaire total perçu (emploi principal, emploi secondaire et autres emplois, y compris rémunérations en nature)
- f) Créer une variable nommée `yeduc` correspondant au nombre d'années d'études suivies

Q.3. Afficher des statistiques de base (moyenne, minimum, maximum) pour les variables créées aux questions Q.2.

Dans la suite de cet exercice, on cherche à expliquer le revenu tiré du travail. Pour cela, on utilise les variables suivantes : `wage`, `sexe`, `age`, `age_sq`, `yeduc`, et l'Etat où vit l'individu, nommée `uf`, que l'on convertira au préalable en variable numérique (commande `egen ... group(...)`).

Q.4. Estimer le modèle par les moindres carrés ordinaires.

Q.5. Lecture du tableau de sortie :

→ Dans la théorie des tests (Neyman et Pearson), le risque de première espèce est communément fixé à 5%. Vous prendrez donc ce seuil comme référence.

- a. Le modèle est-il globalement significatif ? Réécrivez l'hypothèse H_0 , et l'hypothèse alternative. Votre conclusion doit être justifiée au regard de la statistique appropriée et de la probabilité qui lui est rattachée.
- b. Donnez le pouvoir explicatif du modèle.

- c. Etudiez la significativité des variables prises une à une, en réécrivant le test d'hypothèse sous-jacent. Justifiez vos résultats.
- d. Commentez économiquement vos résultats.

Q.6. Déterminez le nombre de modalités de la variable associée à l'Etat de résidence de l'individu. Y a-t-il autant de modalités présentes dans le tableau de résultats ? Pourquoi ?

Q.7. Décrire très précisément la relation entre âge et salaire.

Q.8. Représentez le graphe des résidus en fonction des valeurs estimées de $\hat{y}$. Cela vous permettra d'inspecter les résidus estimés afin de repérer (visuellement) un problème d'hétéroscédasticité. Sous Stata : *rvfplot*
Qu'en déduisez-vous *a priori* ?

Q.9. Réaliser le test d'hétéroscédasticité pour un seuil $\alpha=5\%$ puis conclure. Sous Stata : *estat hettest*

Q.10. Corrigez votre modèle si nécessaire. Comparez le nouveau modèle corrigé avec l'ancien sans correction. Pour cela, utilisez les commandes *estimates store*, et *esttab* (avec l'option « se » pour afficher les erreurs-types).

Q.11. Quel autre problème peut-on suspecter dans notre modèle ? Comment y remédier ? Détaillez.

Q.12. Comparez le nombre d'observations utilisées dans notre modèle à la taille de la base de données. D'où vient cet écart ? Quel problème cela pose-t-il ?

Q.13. Estimez le même modèle que précédemment, mais cette fois pour la probabilité de travailler. Commentez économiquement.

Application 4.

Modèles à variable dépendante dichotomique

On s'intéresse à la participation des femmes mariées au marché du travail. Pour cela, on dispose de données en coupe transversale recueillies aux États-Unis auprès de 753 femmes mariées, âgées de 30 à 60 ans en 1975. Il s'agit d'un extrait de l'enquête PSID (*données issues de Mroz, 1987*). Un bref **dictionnaire des codes** est disponible à la fin de l'exercice.

Vous utiliserez la base de données **womenwork.dta**.

Après avoir ouvert et enregistré un dofile, renseigné le chemin de votre répertoire de travail et ouvert la base de données, on vous demande de procéder aux tâches qui suivent :

Partie 1. Préparation des données

Q.1. Effectuer les opérations suivantes pour préparer les variables qui nous serviront ensuite :

- a) Convertir la variable `inlf` en variable dichotomique prenant les valeurs 0 ou 1 (1 si la femme participe au marché du travail). Labeliser ses modalités comme dans le dictionnaire des codes.
- b) Créer la variable `kidslt6` correspondant au nombre d'enfants de moins de 6 ans
- c) Créer une variable d'âge au carré (`age2`) et d'expérience au carré (`exper2`)
- d) Créer 2 variables dichotomiques relatives au niveau d'éducation :
 - `educ1` si la femme a fait entre 5 et 12 ans d'études
 - `educ2` si la femme a fait plus de 12 ans d'études

Q.2. Afficher des statistiques de base (moyenne, minimum, maximum) pour les variables créées aux questions Q.1.a), b) et d), ainsi que `age`, `nwifeinc`, `educ` et `exper`. Commenter brièvement.

Dans la suite de ce travail, on cherche à expliquer la probabilité pour une femme de participer au marché du travail. Pour cela, on utilise les variables suivantes : `nwifeinc`, `educ2`, `kidslt6`, `kidsge6`, `age`, `age2`, `exper`, `exper2`.

Partie 2. Modèle de probabilité linéaire

Q.3. Pourquoi la variable `educ1` n'est pas incluse dans le modèle ?

Q.4. Estimer le modèle de probabilité linéaire correspondant. Commenter économiquement.

Q.5. Quels sont les potentiels problèmes posés par l'estimation d'un modèle de probabilité linéaire ? En utilisant les commandes suivantes, concluez sur l'existence de ces problèmes :

- `rvfplot`
- `predict yhat, xb`

Q.6. Réaliser le test d'hétéroscédasticité pour un seuil $\alpha=5\%$ puis conclure. La conclusion aurait-elle été la même pour $\alpha=1\%$?

Partie 3. Modèle Probit

Q.7. Estimer le même modèle qu'en question Q.4. en utilisant un Probit et en corrigeant pour l'hétéroscédasticité.

Q.8. A la suite de cette estimation, assurez-vous de la qualité de l'ajustement et de son degré de prédiction à l'aide de la commande suivante : *estat classification*

Q.9. Faites une nouvelle estimation en utilisant le modèle Probit mais en ne conservant que trois variables : `educ2`, `age` et `kidslt6`, et en indiquant bien que la variable `educ2` est une indicatrice.

Q.10. Calculez et interprétez les effets marginaux aux points moyens.

Q.11. Calculez et interprétez les effets marginaux suivants sur la variable dépendante :

- Le fait d'avoir 30 ans, d'avoir un enfant de moins de 6 ans et d'avoir suivi au moins 12 ans d'études.
- Le fait d'avoir 45 ans, d'avoir un enfant de moins de 6 ans et d'avoir suivi au moins 12 ans d'études.
- Le fait d'avoir 60 ans, d'avoir un enfant de moins de 6 ans et d'avoir suivi au moins 12 ans d'études.

Q.12. Calculez les effets marginaux moyens.

Dictionnaire des codes

<i>Nom de la variable</i>	<i>Description et modalités le cas échéant</i>
inlf	Participation au marché du travail 1. Ne participe pas au marché du travail 2. Participe au marché du travail
kids	Nombre d'enfants de moins de 18 ans
kidsge6	Nombre d'enfants ayant entre 6 et 18 ans
age	Age
educ	Nombre d'années de scolarisation
wage	Salaire horaire de la femme
nwifeinc	Montant des autres revenus du ménage (en milliers de dollars)
exper	Expérience professionnelle mesurée par le nombre d'années passées sur le marché du travail depuis les 18 ans de la femme
huseduc	Nombre d'années d'études suivies par le mari
motheduc	Nombre d'années d'études suivies par la mère de la femme
fatheduc	Nombre d'années d'études suivies par le père de la femme

Application 5.

Données de panel

Cet exercice a pour objectif de présenter les commandes de base pour la manipulation des données de panel sous Stata, puis d'estimer une régression à effets fixes et à effets aléatoires. Pour cela, on dispose de la base de données `nlswork.dta`. Ces données longitudinales sont extraites de l'enquête en panel réalisée auprès de femmes ayant entre 14 et 26 ans en 1968, suivies jusqu'en 1988.

Après avoir ouvert et enregistré un dofile, on vous demande de procéder aux opérations suivantes :

Partie 1. Généralités sur les données de panel

Q.1. Ouvrir la base de données en ligne nommée `nlswork.dta` à l'aide de la commande `webuse`

Q.2. Afficher une description des variables présentes dans la base puis des statistiques de base pour chaque variable

Q.3. Identifier les variables correspondant à la dimension individuelle et à la dimension temporelle, puis déclarer le panel dans Stata

Q.4. Afficher la description du panel et commenter de manière précise la sortie Stata

Q.5. Afficher les variables `idcode`, `year`, `age`, `race` et `south` pour les 30ièmes lignes avec la commande `list`

Partie 2. Modèles FE et RE

On s'intéresse aux déterminants du salaire. Pour cela, on veut estimer l'effet de plusieurs variables sur le salaire. Dans la suite du dofile précédent :

Q.6. Créer les variables suivantes : expérience au carré, âge au carré, une indicatrice pour le fait d'avoir la peau noire et le nombre d'années d'études au carré.

Q.7. Estimer le modèle à EF correspondant et sauvegarder les résultats. Commenter.

Q.8. Estimer le modèle à EA correspondant et sauvegarder les résultats. Commenter.

Q.9. Selon vous, vaut-il mieux utiliser un modèle à EF ou à EA ? Justifier.

Q.10. Réaliser le test d'Hausman et conclure.