Université de Grenoble Alpes
(M1 BDA, S2)

ÉCONOMÉTRIE

Travaux

Travail 1

(Cette version : 5 février 2025)

Michal W. Urdanivia [1]

1. Contact : michal.wong-urdanivia@univ-grenoble-alpes.fr, Université de Grenoble Alpes, Faculté d'Économie, GAEL.

Table des matières

1. RÉGRESSION LINÉAIRE SIMPLE

On s'intéresse ici au lien entre une variable $Y \in \mathbb{R}$ et une variable $\in X_{2,i} \in \mathbb{R}$ pour lesquelles on dispose d'observations *i.i.d.*, $\{(Y_i, X_{2,i})\}_{i=1}^n$.

Ces observations vont nous permettre d'estimer les paramètres β_1 et β_2 de l'équation [2]

$$Y_i = \beta_1 + \beta_2 \cdot X_{2,i} + U_i,\ \mathrm{E}(U_i) = 0, \tag{1}$$

où U_i représente l'erreur de notre modèle, soit une variable qui résume les déterminants inobservées pour l'économètre de Y_i.

Pour construire des estimateurs des paramètres nous allons reprendre certaines des hypothèses/-conditions données en cours [3]

Dans un premier temps notons que le vecteur des régresseurs ici est $X_i = \left(1, X_{2,i}\right)^\intercal$ et (1) peut s'écrire aussi [4]

$$Y_i = \beta_1 + \beta_2 \cdot X_{2,i} + U_i = X_i^\intercal \beta + U_i,\ \mathrm{E}(U_i) = 0, \tag{2}$$

où $\beta = \left(\beta_1, \beta_2\right)^\intercal$ est le vecteur des paramètres.

Introduisons/reprenons une condition d'exogénéité de X_i du cours en nous contentant de celle plus *faible* que l'indépendance en moyenne entre U_i et X_i :

$$\mathrm{E}(U_i X_i) = 0 \Leftrightarrow \mathrm{E}\left[(Y_i - X_i^\intercal \beta) X_i\right] = 0, \tag{3}$$

où l'on a employé pour passer de la première à la deuxième expression le fait que d'après (2) $U_i = Y_i - X_i^\intercal \beta$.

On note que dans le cas spécifique qui nous concerne ici avec $X_i = \left(1, X_{2,i}\right)^\intercal$ on a :

$$\begin{cases} \mathrm{E}\left(Y_i - X_i^\intercal \beta\right) = 0, \\ \mathrm{E}\left[(Y_i - X_i^\intercal \beta) X_{2,i}\right] = 0. \end{cases} \tag{4}$$

Dans ce cas les estimateur des moments des paramètres sont obtenus comme solution de [5]

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^n (Y_i - X_i^\intercal \hat{\beta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 \cdot X_{2,i}) = 0, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^n \left[(Y_i - X_i^\intercal \hat{\beta}) X_{2,i}\right] = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \left[(Y_i - \hat{\beta}_1 - \hat{\beta}_2 \cdot X_{2,i}) X_{2,i}\right] = 0 \end{cases} \tag{5}$$

2. Notons que le modèle inclut un premier régresseur constant, égal à 1 par convention, soit en reprenant les notations du cours $X_{1,i} = 1$ avec le paramètre qui lui est associé noté β_1.

3. Remarque : ces hypothèses auront le même contenu mais se présenteront parfois sous forme différente compte tenu du cas particulier de ce travail. Il en va de même des formules qui seront obtenues pour nos estimateurs.

4. $^\intercal$ est le symbol de transposition d'un vecteur/matrice.

5. Pour rappel on utilise à la place des moment théoriques(espérances) des moments empiriques(moyennes) pour définir des estimateurs des paramètres.

où l'on note $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2\right)^\intercal$ le vecteur des estimateurs.

(1) Utiliser le système d'équations précédents pour obtenir les estimateur de β_1 et β_2.

(2) Montrez que ces estimateurs s'écrivent comme fonctions de la moyenne empirique de Y_i, la variance empirique de X_i, et la covariance empirique de Y_i et X_i.

(3) Montrez que ces estimateurs sont en fait ceux des moindres carrés, à savoir il sont solution du problème de minimisation de $\sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 \cdot X_{2,i})^2$ par rapport à β_1 et β_2 :

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta_1,\beta_2} \sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 \cdot X_{2,i})^2$$

2. INFÉRENCE DANS LE MODÈLE DE RÉGRESSION LINÉAIRE NORMAL

Nous allons considérer un modèle où notre intérêt se porte sur la relation entre d'une part le niveau d'études mesuré en année de scolarité depuis l'entrée dans les études initiales jusqu'à la fin des études et l'entrée dans la vie active, et d'autre part le niveau de rémunération. Cela sur une population qui est celle des femmes mariées aux États-Unis.

Notre échantillon est celui employé par Wooldridge (2010) et est issu des données utilisées dans **?**. Il concerne $n = 428$ femmes mariées actives et occupées en 1975.

Les deux principales variables pour notre modèle sont nommées *lwage* et *educ* et correspondent respectivement au logarithme du salaire horaire et au niveau d'études(mesuré comme indiqué plus haut).

Nous disposons aussi d'autres variables telles que l'expérience en nombre d'années, l'âge, le nombre d'enfants, etc. Pour simplifier la présentation nous allons regrouper ces variables dans un vecteur colonne W de dimension $K - 1$(lequel contient aussi un régresseur constant).

Nous observations sont donc un échantillon $\{(W_i, lwage_i, educ_i)\}_{i=1}^{n}$, qu'on suppose aléatoire(ou d'observations i.i.d.).

Notre modélisation commence par supposer que :

$$lwage_i = \beta_1 \cdot educ_i + W_i^\intercal \beta_2 + U_i \tag{6}$$

où β_1 est à paramètre scalaire inconnu, β_2 un vecteur de paramètres inconnus de même dimension que W_i, et U_i désignera l'erreur du modèle pour lequel on supposera que

$$\mathrm{E}(U_i | educ_i, W_i) = 0 \tag{7}$$

(1) Qualifiez le type de données utilisées(panel, en coupe, séries temporelles ?)

(2) Interprétez le coefficient β_1 en supposant (6) et (7).

(3) Considérez (6) comme l'équation d'une *régression partitionnée* et indiquez une procédure pour obtenir l'estimateur des MCO du paramètre β_1 qui n'utilise pas l'estimation par MCO dans la régression de $lwage_i$ sur $educ_i$ et W_i mais donne le même paramètre estimé.

(4) On note $\hat{\beta}_1$ l'estimateur des MCO de β_1 et $\hat{\mathrm{Var}}(\hat{\beta}_1|\mathbf{X})$ sa variance estimée($\mathbf{X}$ est la matrice des régresseurs associé à (6)). On obtient $\hat{\beta}_1 = 0.108$, et une variance estimée telle que $\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_1|\mathbf{X})} = 0.014$(on se place ici dans le cadre le plus courant où la variance des erreurs supposés homoscédastiques est inconnue).

Nous allons supposer que les conditions d'un *modèle de régression linéaire normal* sont satisfaites. Donnez un intervalle de confiance de niveau $\alpha = 0.05$ pour β_1 par(où 1.96 est le quantile $t_{n-K,1-\alpha/2}$ de la loi de Student). Interprétez cet intervalle.

(5) On cherche à tester l'hypothèse que $\beta_1 = 0$. Donnez une statistique de test(et sa valeur ici).

(6) L'hypothèse est-elle rejetée/acceptée pour un test de niveau $\alpha = 0.05$(ici aussi le quantile $t_{n-K,1-\alpha/2}$ de la loi de Student est 1.96) ? Justifiez.

Références

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. The MIT Press.