

Exercices pour le cours Econométrie et évaluations statistiques (Ch. V).

Propriétés statistiques des MCO II: Quelques exemples – Efficacité et normalité

Exercice 1. Considérez une régression dans laquelle il n'y a qu'une constante et aucune variables explicatives. Étant donné un échantillon $\{y_i : i = 1, \dots, N\}$, le problème de minimisation est:

$$\min_{a_0} \sum_{i=1}^N (y_i - a_0)^2$$

1. Soit $\hat{\alpha}_0$ la solution de cette minimisation. Montrez que

$$\hat{\alpha}_0 = \frac{1}{N} \sum_{i=1}^N y_i,$$

c'est-à-dire que la moyenne de l'échantillon minimise la somme des carrés des résidus.

2. Notez que $y_i = \alpha_0 + u_i$ et supposez que les hypothèses H2-H5 sont satisfaites (dont l'homoscédasticité). Calculez la variance de $\hat{\alpha}_0$.
3. Définissez les résidus $\hat{u}_i = y_i - \bar{y}$. Montrez que ces résidus somment sur zéro.
4. Utilisez ce résultat avec le Théorème 2 pour donner une interprétation à la variance de l'estimateur du paramètre de pente dans la modèle de régression simple:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

est égal à

$$\text{Var}(\hat{\beta}_1 | x_1, \dots, x_N) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}.$$

Exercice 2. L'équation suivante a été estimée:

$$\widehat{\text{sleep}} = 3840,83 - 0,16 \text{ totwrk} - 11,71 \text{ educ} \\ (235,11) \quad (0,02) \quad (5,86) \\ - 8,70 \text{ age} + 0,13 \text{ age}^2 + 87,75 \text{ male} \\ (11,21) \quad (0,13) \quad (34,33)$$

$$N = 706, \quad R^2 = 0,123$$

où sleep désigne le nombre total de minutes passées à dormir par semaine, totwrk le nombre total de minutes passées à travailler par semaine, educ l'éducation en années, l'âge en années, et male est une variable binaire qui prend la valeur 1 pour les hommes et 0 pour les femmes.

1. (a) A quel âge, en moyenne, les individus dorment-ils le moins?
- (b) Comment interprétez-vous l'estimation du paramètre associé à la variable male?
- (c) Soit une variable female qui prend la valeur 1 pour les femmes et 0 pour les hommes. Quelles seraient les conséquences de l'introduction de cette variable additionnelle dans la régression?

Exercice 3. Soit la base de données NBASAL.xls qui contient des informations sur les salaires et des statistiques de carrière pour 269 joueurs de la National Basketball Association (NBA).

1. Estimez un modèle reliant le nombre de points par match (points) aux années dans la ligue (exper), à l'âge, et aux années jouées à l'université (coll). Incluez un terme quadratique pour l'expérience. Rapportez les résultats de la manière habituelle.
2. En maintenant l'âge et les années jouées à l'université constants, pour quelle valeur de l'expérience une année supplémentaire d'expérience aura-t-elle un effet négatif sur le nombre de points par match? Est-ce que cela a du sens?

3. Pourquoi pensez-vous que coll a un coefficient négatif? [Indice: les joueurs de la NBA peuvent être repêchés avant de terminer leur carrière universitaire et même directement à la sortie du lycée.]

Exercice 4. Expliquez la différence entre “efficacité” et “biais” d’un estimateur. Peut-on avoir un estimateur biaisé mais plus efficace (donc avec une plus petite variance) qu’un estimateur non biaisé ? Donnez un exemple (conceptuel ou empirique).

Exercice 5. Un économiste cherche à mesurer l’impact de la taille d’une entreprise sur le salaire moyen de ses employés. Il estime, à partir de 250 observations, le modèle suivant :

$$\text{wage}_i = \beta_0 + \beta_1 \log(\text{employees}_i) + u_i$$

où wage_i est le salaire moyen annuel (en milliers d’euros) dans l’entreprise i , employees_i est le nombre total de salariés de l’entreprise. L’estimation par les MCO donne:

$$\begin{aligned} \widehat{\text{wage}} &= 22,5 + 3,8 \log(\text{employees}) \\ &\quad (1,9) \quad (0,6) \\ R^2 &= 0,45 \end{aligned}$$

1. Interprétez le coefficient β_1 en tenant compte de la forme lin-log. Donnez la valeur chiffrée en milliers d’euros de l’effet sur le salaire moyen d’une augmentation de 20% du nombre d’employés. Considérez une valeur approximative et une valeur exacte.
2. Si on ajoutait une variable comme le secteur d’activité ou le niveau d’études moyen des travailleurs, pensez-vous que le coefficient sur $\ln(\text{employees})$ changerait ? Pourquoi ?