

Exercices d'économétrie

TD 4. Régression multiple: inférence.

Janvier, 2020

Exercice 4.1

Le modèle suivant peut être utilisé pour étudier si les dépenses de campagne influent sur les résultats des élections :

$$\text{voteA}_i = \beta_0 + \beta_1 \log(\text{expendA}_i) + \beta_2 \log(\text{expendB}_i) + \beta_3 \text{prtystrA}_i + u_i$$

où voteA est le pourcentage de vote obtenu par le candidat A , expendA et expendB les dépenses de campagne des candidats A et B , et prtystrA une mesure de l'importance du parti pour le candidat A (le pourcentage de votes en faveur du parti du candidat A lors des élections les plus récentes).

- i Quelle est l'interprétation du paramètre β_1 ?
- ii En fonction des paramètres du modèle, spécifiez l'hypothèse nulle selon laquelle l'impact sur le vote d'une augmentation de 1% des dépenses de A est compensée par une augmentation de 1% des dépenses de B .
- iii Estimez le modèle en utilisant la base de données `vote1`. Les dépenses du candidat A influencent-elles le résultat ? Qu'en est-il des dépenses du candidat B ? Pouvez-vous utiliser ces résultats pour tester l'hypothèse de la question (ii) ?
- iv Estimez un modèle donnant directement la statistique de Student adéquate pour tester l'hypothèse décrite en question (ii). Que concluez-vous ?

Exercice 4.2

On utilise la base de données `hprice1` pour estimer le modèle :

$$\log(\text{price}_i) = \beta_0 + \beta_1 \text{sqrft}_i + \beta_2 \text{bdrms}_i + u_i$$

où price est le prix d'une maison (en milliers de dollars), sqrft mesure sa superficie (en pieds carrés), et bdrms représente le nombre de chambres.

- i On s'intéresse à l'estimation et l'obtention d'un intervalle de confiance pour le taux de variation en pourcentage de la variable price lorsqu'une pièce de 150 pieds carrés est ajoutée à la configuration existante d'une maison donnée. Sous forme décimale, ceci s'écrit $\theta = 150\beta_1 + \beta_2$. Estimez θ .
- ii Écrivez β_2 en fonction de β_1 et θ et introduisez ce paramètre dans l'équation expliquant $\log(\text{price})$.
- iii Utilisez les éléments de la question (ii) afin d'obtenir un écart-type estimé pour $\hat{\theta}$, puis calculez à partir de cet écart-type un intervalle de confiance à 95%.

Exercice 4.3

On utilise la base de données `mlb1` pour expliquer les salaires des joueurs de baseball de la ligue professionnelle américaine à l'aide du modèle suivant :

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \text{years}_i + \beta_2 \text{gamesyr}_i + \beta_3 \text{bavg}_i + \beta_4 \text{hrunsyr}_i + \beta_5 \text{rbisyr}_i + u_i$$

où salary est le salaire total en 1993, years le nombre d'années passées au sein de la ligue, gamesyr le nombre moyen de rencontres disputées dans l'année, bavg la « moyenne à la batte » sur l'ensemble de la carrière du joueur, hrunsyr le nombre de « *home runs* » par an, et rbisyr le nombre de points produits par un frappeur au cours de l'année.

- i Estimez le modèle avec et sans la variable rbisyr . Qu'advient-il de la significativité statistique de hrunsyr ? Qu'en est-il de la taille du coefficient de hrunsyr ?

- ii Ajoutez les variables `runsy` (nombre de points par an), `fldperc` (pourcentage de mise en service), et `sbasesyr` (base « volées » par an) au second modèle de régression de la question (i). Lequel de ces facteurs est-il individuellement statistiquement significatif?
- iii Dans le modèle de la question (ii), testez la significativité jointe de `bavg`, `fldperc`, et `sbasesyr`.

Exercice 4.4

On utilise la base de données `discrim` qui contient des données de prix pour différents produits vendus dans des enseignes de restauration rapide.

- i Estimez le modèle suivant par les MCO :

$$\log(\text{psoda}_i) = \beta_0 + \beta_1 \text{prpblck}_i + \beta_2 \log(\text{income}_i) + \beta_3 \text{prppov}_i + u_i$$

où `psoda` est le prix du soda, `prpblck` la proportion locale de personnes d'origine afro-américaine, `income` le revenu local moyen, et `prppov` la proportion locale de personnes vivant sous le niveau de pauvreté. Le paramètre estimé $\hat{\beta}_1$ est-il statistiquement différent de zéro au seuil de 5% contre l'hypothèse alternative bilatérale? Qu'en est-il au seuil de 1%?

- ii Quelle est la corrélation entre `log(income)` et `prppov`? Les variables apparaissent-elles chacune statistiquement significative?
- iii Reprenez le modèle de la question (i) et ajoutez-y la variable `log(hseval)` (la valeur du logement). Interprétez son coefficient et indiquez la p-valeur relative au test bilatéral pour $H_0 : \beta_4 = 0$.
- iv Dans la régression décrite en question (iii), que se passe-t-il pour la significativité statistique individuelle des variables `log(income)` et `prppov`? Ces variables sont-elles conjointement significatives? Qu'advient-il de vos précédentes réponses?
- v Compte tenu des résultats des régressions précédentes, laquelle seriez-vous prêt à reporter avec le plus de confiance afin d'analyser si la composition ethnique d'une localité influence les prix des fast-foods?