Indications sur le modèle de régression linéaire

Catherine Bruneau

2004-2005

1 Recherche de la droite des moindres carrés

Principe de l'ajustement d'une droite à un nuage de points

Qualité de l'ajustement, le coefficient R^2, la décomposition de la variance.

Voir un manuel; par exemple, chapitre 1 du livre de B. Dormont, Introduction à l'économétrie

2 Le modèle de régression liénaire simple

Le modèle de régression linéaire simple:

$$Y_i = \alpha + \beta X_i + \varepsilon_i; 1 \leq i \leq n$$

Les hypothèses des Mondres Carrés Ordinaires (MCO).

A partir de l'écriture du modèle linéaire,

$$\begin{pmatrix} Y_1 \\ . \\ \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_1 \\ . & . \\ \\ 1 & X_n \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ \\ \varepsilon_n \end{pmatrix}$$

les hypothèses se formulent de la manière suivante:

$$\begin{aligned} i) E(\overrightarrow{\varepsilon}/\mathcal{X}) &= \overrightarrow{0} \\ ii) Var(\overrightarrow{\varepsilon}/\mathcal{X}) &= \sigma^2 Id \\ & iii) \mathcal{X}'\mathcal{X} \text{ est inversible} \\ & iv) \frac{1}{N}\mathcal{X}'\mathcal{X} \underset{n\to\infty}{\to} Q \end{aligned}$$

où Id désigne la matrice identité de taille n.

Les hypothèses:

i) les résidus sont centrés, pris conditionnellement à la variable explicative.

ii) la variance de ε_i, calculée, conditionnellement à la valeur de X_i est constante ($= \sigma^2$) quel que soit l'individu i (hypothèse d'homoscédasticité).

De plus, les résidus correspondant à deux individus différents i et j ne sont pas corrélés, conditionnellement aux valeurs de X_i et X_j.

iii) les colonnes de $\mathcal{X}$ ne sont pas colinéaires: la variable explicative X ne doit pas être constante sur les individus.

iv) La matrice $\mathcal{X}'\mathcal{X}$ n'est pas explosive lorsque n tend vers l'infini

On demande aussi que la suite $\frac{1}{N}\mathcal{X}'\mathcal{X}$ soit convergente lorsque n tend vers l'infini.

Voir:

- les résultats de l'estimation (théorème de Gauss Markov).
- les propriétés des estimateurs des MCO $\widehat{\alpha}$ et β
- les principaux tests (test de significativité, test d'analyse de la variance, test d'une contrainte linéaire sur les paramètres α et β)
- intervalles de confiance sur les paramètres
- prévision et intervalle de confiance sur la valeur prévue

Voir chapitre 1 du livre de B. Dormont

3 Modèle de régression linéaire multiple

Généralisation du cas précédent.

Le modèle de régression s'écrit:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_K X_{iK} + \varepsilon_i; 1 \leq i \leq n$$

Hypothèses des MCO: même écriture des hypothèses avec une matrice $\mathcal{X}$ plus étoffée:

$$\mathcal{X} = \begin{bmatrix} 1 & \mathcal{X}_{11} & . & \mathcal{X}_{1K} \\ . & . & . & . \\ . & . & . & . \\ 1 & \mathcal{X}_{n1} & . & \mathcal{X}_{nK} \end{bmatrix}$$

Remarque: colonne de 1 si le modèle comporte une constante.

Résultats de l'estimation des MCO: estimateurs des MCO des paramètres $\beta_k, 0 \leq k \leq K$ de $\sigma^2 = Var(\varepsilon_i/\mathcal{X})$:

$$\begin{aligned} \overrightarrow{\widehat{\beta}} &= (\widehat{\beta_0}, ..., \widehat{\beta_K})' = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\overrightarrow{Y} \\ \widehat{\sigma^2} &= \frac{1}{n-K+1}\sum_{i=1}^{n} e_i^2 \\ \text{où } e_i &= Y_i - \widehat{\beta_0} - \sum_{k=1}^{K} \widehat{\beta_k} X_{ik} \end{aligned}$$

Voir:

- les propriétés de ces estimateurs.
- les ntervalles de confiance sur les paramètres estimés.
- les principaux tests sur les paramètres $\beta_k, 0 \leq k \leq K$.
- la prévision.

Voir chapitre 2 et chapitre 3 du livre de B. Dormont.

4 Remise en question des hypothèses des MCO

Les hypothèses qui sont remises en question concernent l'homoscédasticité et/ou l'auto-corrélation des résidus.

Les hypothèses(des moindres carrés généralisés) deviennent:

$$\begin{aligned} i)E(\overrightarrow{\varepsilon}/\mathcal{X}) &= \overrightarrow{0} \\ ii)Var(\overrightarrow{\varepsilon}/\mathcal{X}) &= \sigma^2\Omega \\ & \quad iii)\mathcal{X}'\mathcal{X} \text{ est inversible} \\ & \quad iv)\frac{1}{N}\mathcal{X}'\mathcal{X} \underset{n\to\infty}{\to} Q \end{aligned}$$

où Ω est une matrice différente de la matrice identité,

soit parce qu'il y a hétéroscédasticité:

$$\exists i \neq j/\ Var(\varepsilon_i/\mathcal{X}) \neq Var(\varepsilon_j/\mathcal{X})$$

soit parce qu'il existe les résidus sont auto-corrélés.

$$\exists i \neq j/\ cov(\varepsilon_i, \varepsilon_j/\mathcal{X}) \neq 0$$

soit parce qu'il ya hétéroscédasticité et auto-corrélation des résidus.

Deux façons de résoudre le problème:

1) on transforme les données (shpéricisation du modèle) et on se ramène au cas d'application des MCO

2) on continue à estimer selon la méthode des MCO, mais il faut modifier les statistiques lorsqu'on effectue des tests sur les paramètres β_k ou lorsqu'on calcule des intervalles de confiance sur ces paramètres. En effet, la précision des estimateurs est affectée par l'hétéoscédasticité ou l'auto-corrélation des résidus

4.1 Exemple de sphéricisation, dans le cas d'hétéroscédasticité

On étudie le profit P_i d'entreprises i en fonction d'un certain nombre de facteurs explicatifs et on constate que les résidus n'ont pas la même variance dans la classe des grandes entreprises et dans la classe des petites entreprises.

$$P_i = \beta_0\ + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i; 1 \leq i \leq n$$

On suppose que la variance des résidus est telle que:

$$Var(\varepsilon_i/\mathcal{X}) = \lambda \mathcal{T}_i$$

où $\mathcal{T}_i$ désigne la taille de l'entreprise.

On transforme alors les données en divisant les deux membres des équations précédentes par $\sqrt{T_i}$ (pour l'équation correspondant à l'entreprise i):

$$\frac{P_i}{\sqrt{T_i}} = \frac{1}{\sqrt{T_i}}\beta_0\ + \sum_{k=1}^{K} \beta_k \frac{X_{ik}}{\sqrt{T_i}} + \frac{\varepsilon_i}{\sqrt{T_i}}; 1 \leq I \leq n$$

soit (modèle sphéricisé):

Il est facile de vérifier que ce modèle satisfait les hypothèses des MCO. On estime alors les différents parmètres par la méthdoe MCO à partir du modèle transformé (remarque: ce modèle n'a plus de constante, sauf si l'une des variables explicatives $X_{.k}$ est la racine carrée de la taille).

En pratique, on ne sait pas comment spécifier $\sigma_i^2 = Var(\varepsilon_i/\mathcal{X})$. On adopte une explication linéaire en fonction d'un certain nombre de facteurs:

$$\sigma_i^2 = \alpha_0 + \sum_{l=1}^{L} \alpha_l Z_{il}$$

ce que l'on traduit empiriquement par:

$$e_i^2 = \alpha_0 + \sum_{l=1}^{L} \alpha_l Z_{il} + u_i$$

où e_i désigne l'estimation du résidu ε_i de la première régression.

Remarque: e_i^2 est un estimateur convergent de ε_i^2. Par conséquent $E(e_i^2) \simeq E(\varepsilon_i^2) = \sigma_i^2$, et par conséquent:

$$e_i^2 = \sigma_i^2 + u_i$$

avec u_i centré et homogène à un résidu si n et grand.

En pratique, on estime donc les paramètres α_l correspondant aux différents facteurs $Z_{.l}$ explicatifs de l'hétéroscédasticité. On fait une estimation par MCO et on estime σ_i^2 par $\widehat{\sigma_i^2} = \widehat{\alpha_0} + \sum_{l=1}^{L} \widehat{\alpha_l} Z_{il}$. On peut ensuite transformer les données en divisant par les variables P_i,$1, X_{.k}$, $1 \leq k \leq KM$, par $\widehat{\sigma_i}$ et estimer par MCO les paramètres β_k . Il faut cependant qu'il y ait suffisamment d'observations (n grand).

La deuxième façon de traiter le problème consiste à appppliquer les MCO: on obtient ainsi des estimateurs convergents de paramètres β_k même en présence d'hétéroscédasticité et/ou d'auto-corrélation des résidus. Mais il faut appliquer des corrections pour mener une inférence correcte.

4.2 Estimation des MCO et correction de White ou HAC (correction del'hétéroscédasticité et de l'auto-corrélation)

1) les estimateurs $\widehat{\beta_k}$ des MCO des paramètres β_k restent des estimateurs sans biais et convergents

2) par contre leur précision est mal estimée si on utilise la formule des MCO:

$$Var(\overrightarrow{\beta}/\mathcal{X}) = \sigma^2(\mathcal{X}'\mathcal{X})^{-1}$$

La variance de $\overrightarrow{\beta}$ est en effet donnée par:

$$Var(\overrightarrow{\beta}/\mathcal{X}) = \sigma^2 \left(\mathcal{X}'\mathcal{X}\right)^{-1} \left(\mathcal{X}'\Omega\mathcal{X}\right) \left(\mathcal{X}'\mathcal{X}\right)^{-1}$$

si $Var(\overrightarrow{\varepsilon}/\mathcal{X}) = \sigma^2\Omega$.

White a proposé une estimation de la matrice de variance $Var(\overrightarrow{\beta}/\mathcal{X})$ en présence d'hétéroscédasticité (acceptable lorsque n est suffisamment grand).

$$\begin{array}{rcl} & & n\left(\mathcal{X}'\mathcal{X}\right)^{-1} S_0 \left(\mathcal{X}'\mathcal{X}\right)^{-1} \\ \text{où } S_0 & = & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} e_i^2 x_i' x_i \\ & & \text{avec } x_i \ i-^{ième} \text{ ligne de } \mathcal{X} \end{array}$$

Lorsqu'il existe des corrélations entre les résidus, il faut aussi corriger l'estimation des MCO de la variance $Var(\overrightarrow{\beta}/\mathcal{X})$.

D'une manière générale, on considère l'inférence sur le modèle après correction pour hétéroscédasticité et auto-corrélation (HAC: heteroskedasticity and auto-correlation correction).

4.3 Test d'hétéroscédasticité

Il existe différents types de test d'hétéroscédasticité. Voir manuel

On peut citer le test deWhite utilisé pour tester l'hypothèse nulle:

$$H_0 : \alpha_1 = ... = \alpha_Q$$

contre $H_1 : \exists k \neq l / \alpha_l \neq \alpha_k$.

Le test est un test d'analyse de la variance associé à la régression de e_i^2 sur Q régresseurs choisi parmi les variables X_{ik},X_{ik}^2,$X_{ik}X_{ik'}, k \neq k'$ pour $1 \leq k \leq K, 1 \leq k' \leq K$.

$$e_i^2 = \alpha_0 + \alpha_1 X_{i1} + ... + \alpha_K X_{iK} + \alpha_{K+1} X_{i1}^2 + ... + v_i$$

On peut réaliser le test en calculant une statistique de Fisher ou bien en considérant le coefficient R^2et plus précisément nR^2 qui suit un chi-deux à $n - Q + 1$ dgrés de liberté, si l'hypothèse nulle est vraie.

Avertissement: si on considère des régressions faisant intervenir des séries chronologiques (des variables qui évoluent au cours du temps), il faut s'assurer que les séries sont stationnaires pour pouvoir appliquer les régles d'inférence qui ont été décrites ci-dessus.

Une série $(X_t)_t$ est stationnaire si son auto-corrélogramme décroit rapidement, en définissant l'auto-corrélogramme comme la représenttaion de la fonction:

$$h \rightarrow \text{corrélation}(X_t, X_{t+h})$$

Le vérifier en traçant l'auto-corrélogramme.

Indications bibliographiques

Introduction à l'économétrie, Brigitte Dormont, édition Montchrestien, 1999.