CHOIX DE MODELES

1 Comment construire un modèle ?

En général, lorsque l'on fait une étude économétrique, on se base sur un modèle économique que l'on souhaite tester sur nos données, on introduit donc les variables du modèle théorique. Si ce n'est pas le cas, on dispose d'un panier de variables dans lequel on veut chercher le meilleur modèle. Pour cela on fait tous les modèles possibles des variables prises deux à deux puis trois à trois .... et on regarde le modèle qui a le plus petit s (s^2 étant l'estimation de σ^2). matériellement il faut pour cela un logiciel qui fasse ce travail un peu long. En pratique, il est bon de regarder non pas le meilleur s mais les deux ou trois meilleurs et de choisir entre ces par un raisonnement économique. On construit ainsi le meilleur modèle **statique**. La comparaison ne peut se faire que si l'on a la même variable endogène Y (ou la même transformation de Y) et si l'on a exactement le même échantillon, ces remarques sont très importantes, on en verra un exemple dans la prochaine section.
Nous verrons dans le chapitre sur les modèles dynamiques que le nombre de retards sur les variables est choisi à l'aide de critères comme AIC, AICC ou BIC car les variables sont ordonnées.
En pratique on commence par un modèle statique, on chosit les variables par le critère du s puis sur ces variables on met des retards avec les critères AIC ..... On construira alors un modèle **dynamique.**
On prend souvent l'habitude d'effectuer les tests classiques sur le modèles statique, pour remarquer qu'ensuite beaucoup des problèmes de tests constatés disparaissent quand on ajoute des retards aux variables endogènes et exogènes. Nous verrons dans le chapitre suivant que les mauvais tests ne doivent pas être traités dans le modèle statique, mais que l'on doit attendre la construction du modèle dynamique qui règle souvent beaucoup de problèmes comme l'autocorrélation et l'hétéroscédasticité même si elle ajoute des problèmes de colinéarité (voir dernier chapitre). Comme nous le verrons également, le fait de mettre des retards sur la variable endogène permet de prendre en compte indirectement des variables "oubliées".
Il existe aussi des techniques comme la régression pas à pas (stepwise) qui consiste à ajouter une à une des variables en prenant la variable la plus corrélée avec y puis en ajoutant la deuxième variable la plus corrélée ... et en s'arrêtant quand le s ne diminue presque plus ou augmente. Mais cette méthode est moins bonne que le choix parmi tous les modèles à cause des problèmes de colinéarité.

2 Choix entre un modèle linéaire et un modèle en log

Sans faire toute la démonstration nous allons choisir entre un modèle statique linéaire et un modèle en log contenant les mêmes variables.
Soit un modèle

$$\text{Modèle linéaire } y = a_0 + a_1 x_1 + a_2 x_2 + \epsilon_0$$

Ce modèle est linéaire, si on hésite entre ce modèle et un modèle en log avec les mêmes variables, ces variables explicatives doivent être strictement positives.

$$\text{Modèle en log: } \text{Log}(y) = b_0 + b_1 Log(x_{1)} + a_2 Log(x_2) + \epsilon_1$$

Nous ne pouvons comparer les s car ils n'ont pas les mêmes unités, les erreurs du deuxième modèle sont en Log des unités du premier.
On utilise la transformation de Box-Cox qui consiste à tranformer toutes les variables de la façon suivante:
les variables x sont transformées en variables

$$x^{(\lambda)} = \frac{x^\lambda - 1}{\lambda}$$

On remarque que

- si $\lambda = 1$ on retrouve le modèle linéaire à une constante près.
- si $\lambda = 0$ on trouve le Log(x) dans le modèle en Log

Si on calcule le maximum de vraisemblance pour les variables $z^{(\lambda)}$ on obtient

$$Log(\text{fonction de vraisemblance}) = cte - \frac{n}{2}(2Log(s) - (\lambda - 1)\overline{Log(y)})$$

$$\max(Log(f) \Longleftrightarrow \min(Log(s - (\lambda - 1)\overline{Log(y)} \Longleftrightarrow \min se^{-(\lambda-1)\overline{(Log(y)}}$$

où $\overline{Log(y)}$ est la moyenne des Log des y_i et dans lequel on estime σ^2 par SRC/(n-k)

- Dans le modèle linéaire pour $\lambda = 1$ le maximum de vraisemblance consiste à min s_0 (s_0 étant le s du modèle linéaire)
- Dans le modèle en Log pour $\lambda = 0$ le maximum de vraisemblance consiste à min $s_1 e^{\overline{(Log(y)}}$ (s_1 étant le s du modèle en Log)
- Conclusion: on préfère le modèle linéaire si $s_0 < s_1 e^{\overline{(Log(y)}}$ et le modèle en Log si $s_1 e^{\overline{(Log(y)}} < s_0$