FORMATION « MODÈLES DE PRÉVISION »

Généralités sur les modèles de prévision

INTERVENANTS

Morgane Glotain
Alain Quartier-la-Tente

29-31 mars 2017

Introduction (1/2)

Objectif : avoir des modèles de prévision simples, facilement estimables et interprétables économiquement

$\rightarrow$ pas de boîte noire (on comprend d'où viennent les prévisions des modèles, les effets de chaque variable) et facilement explicable (intérêt dans la diffusion)

Méthode populaire dans les INS et les banques centrales : *bridge equations* (**étalonnage**)

Introduction (2/2)

Modèles retenus : modèles de régression linéaire estimables par moindres carrés ordinaires (MCO)

$$\underbrace{y_t}_{\substack{\text{Taux de croissance}\\ \text{de la variable à prévoir}}} = \beta_0 + \sum_{n=1}^{N} \underbrace{\beta_n}_{\substack{\text{Élasticité}}} \overbrace{x_{n,t}}^{\substack{\text{Variable explicative}\\ \text{(soldes, ind. quanti, etc.)}}} + \underbrace{\sum_{m=1}^{4} \zeta_m y_{t-m}}_{\substack{\text{Retards de la variable}\\ \text{endogène}}} + \underbrace{\varepsilon_t}_{\substack{\text{Résidus}}}$$

Introduction (2/2)

Modèles retenus : modèles de régression linéaire estimables par moindres carrés ordinaires (MCO)

$$\underbrace{y_t}_{\substack{\text{Taux de croissance}\\ \text{de la variable à prévoir}}} = \beta_0 + \sum_{n=1}^{N} \underbrace{\beta_n}_{\text{Élasticité}} \overbrace{x_{n,t}}^{\substack{\text{Variable explicative}\\ \text{(soldes, ind. quanti, etc.)}}} + \underbrace{\sum_{m=1}^{4} \zeta_m y_{t-m}}_{\substack{\text{Retards de la variable}\\ \text{endogène}}} + \underbrace{\varepsilon_t}_{\text{Résidus}}$$

Pour que les estimations par MCO soient **sans biais**, **consistantes** (convergent vers la vraie valeur) et **efficaces** il faut faire plusieurs hypothèses sur :

- les variables explicatives : stationnaires et exogènes
- les résidus : décorrélés, homoscédastiques et suivent une loi normale

Sommaire

La stationnarité qu'est-ce que c'est ?

Définition :

Idée : (X_t) est **stationnaire** si

- la moyenne de X_t est constante dans le temps (pas de rupture de série en niveau, pas de tendance)
- la variance de X_t est constante dans le temps
- $\text{Cov}(X_{t-k}, X_t)$ ne dépend que de k

$\rightarrow$ On regarde en taux de croissance les variables issues de données quantitatives (IPI, ICA, comptes trimestriels, *etc.*)

Pourquoi c'est important ?

Notion importante dans le cadre des prévisions. Si les variables ne sont pas stationnaires :

- la prévision est impossible (le passé n'apporte pas d'information pour la prévision)
- les étalonnages ne sont plus valides (pas de convergence des estimateurs, lois limites non valides)
- régression fallacieuse : régression linéaire fait apparaître des liens entre des variables alors que ce n'est pas le cas (ex. : nombre de prix Nobel par pays et consommation annuelle de chocolat par habitant)

L'exogénéité qu'est-ce que c'est ? (1/2)

Définitions : exogénéité et endogénéité

Dans le modèle $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ on dit que :

- x_t est **fortement exogène** si $\forall t, t' : \mathbb{E}[x_t \varepsilon_{t'}] = 0$
- x_t est **faiblement exogène** si $\forall t : \mathbb{E}[x_t \varepsilon_t] = 0$
- x_t est **endogène** dans le cas contraire

L'endogénéité peut provenir : d'une erreur de mesure, d'autocorrélation des résidus, d'une variable omise, d'une causalité simultanée entre la variable explicative et la variable à prévoir (ex : équation offre-demande, le prix est endogène pour prévoir la demande à l'équilibre)

L'exogénéité qu'est-ce que c'est ? (2/2)

On suppose que dans les étalonnages des agrégats macroéconomiques (comptes trimestriels, IPI, *etc.*) les **soldes sont fortement exogènes** : les réponses des entreprises ne décrivent que leur situation propre et ne dépendent pas des chiffres publiés par la comptabilité nationale

L'exogénéité qu'est-ce que c'est ? (2/2)

On suppose que dans les étalonnages des agrégats macroéconomiques (comptes trimestriels, IPI, *etc.*) les **soldes sont fortement exogènes** : les réponses des entreprises ne décrivent que leur situation propre et ne dépendent pas des chiffres publiés par la comptabilité nationale

Exemple :

Dans la régression $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$ la variable y_{t-1} n'est pas fortement exogène :

$$\mathbb{E}\left[y_{(t-1)+1}\varepsilon_t\right] = \mathbb{E}\left[y_t\varepsilon_t\right] \neq 0$$

Pourquoi c'est important ?

Les hypothèses d'exogénéité sont importantes :

- si une variable explicative est endogène : les estimations sont biaisées et ne sont pas consistantes (ne convergent pas)
- si une variable explicative est faiblement exogène : les estimations sont consistantes (sans biais asymptotiquement, *i.e.* : l'erreur diminue quand la taille d'échantillon augmente) mais peuvent être biaisées
- si une variable explicative est fortement exogène : les estimations sont non biaisées et consistantes

Pourquoi c'est important ?

Les hypothèses d'exogénéité sont importantes :

- si une variable explicative est endogène : les estimations sont biaisées et ne sont pas consistantes (ne convergent pas)
- si une variable explicative est faiblement exogène : les estimations sont consistantes (sans biais asymptotiquement, *i.e.* : l'erreur diminue quand la taille d'échantillon augmente) mais peuvent être biaisées
- si une variable explicative est fortement exogène : les estimations sont non biaisées et consistantes

Exemple dans le cas d'un $AR(1)$: $\forall t \in [\![2\,,\,T]\!]\ :\ y_t = \beta y_{t-1} + \varepsilon_t$

$$\hat{\beta} = \beta + \frac{\frac{1}{T-1}\sum_{t=1}^{T-1} \varepsilon_{t+1} y_t}{\frac{1}{T-1}\sum_{t=1}^{T-1} y_t^2}$$

$$\underset{T\to+\infty}{\text{plim}}\ \hat{\beta} = \beta \text{ si } \mathbb{E}[\varepsilon_{t+1} y_t] = \mathbb{E}[\varepsilon_t y_{t-1}] = 0$$

Sommaire

Les hypothèses sur les résidus

Pour que les propriétés précédentes soient valides il faut également des propriétés sur les résidus ε_t :

- ils doivent être **décorrélés** : $\forall t \neq t' : \text{Cov}(\varepsilon_t, \varepsilon_{t'}) = 0$ (dans le cas contraire on parle d'**autocorrélation**)
- ils doivent être **homoscédastiques** : $\forall t, t' : \mathbb{V}[\varepsilon_t] = \mathbb{V}[\varepsilon_{t'}]$ (variance constante dans le temps ; dans le cas contraire on parle d'**hétéroscédasticité**)
- ils doivent suivre une **loi normale**

Dans ce cas, et si les variables explicatives sont exogènes, les estimateurs par MCO sont consistants (convergent vers la vraie valeur) et efficaces (de variance minimale) : on parle d'estimateur *BLUE*

Autocorrélation : sources et conséquences (1/2)

Souvent présente dans les séries temporelles, l'autocorrélation des résidus peut provenir :

- d'erreurs de mesure : si les données sont interpolées toujours à la même date, un biais systématique peut être observé
- problème de variable omise : il manque une variable explicative importante
- mauvaise spécification : par exemple dans le cas d'une équation non linéaire mais polynomiale
- d'un lissage artificiel des données trimestrielles sur données annuelles

Autocorrélation : sources et conséquences (1/2)

En cas d'autocorrélation des résidus :

- pas de biais supplémentaire dans les estimations des coefficients
- estimations ne sont plus efficaces (de variance minimale), même asymptotiquement : variances et erreurs de prévision sous-estimées, tests de Student invalidés
- les estimations peuvent ne pas être consistantes (ex : si la variable omise est un retard de l'endogène)
- invalidité des tests d'hétéroscédasticité (supposent résidus décorrélés)


l'autocorrélation peut souvent être corrigée avec des retards de la variable endogène

Sources d'hétéroscédasticité et conséquences

Dans le cadre de séries temporelles, l'hétéroscédasticité provient souvent de la présence de points atypiques

En présence hétéroscédasticité :

- estimations ne sont plus efficaces (de variance minimale), même asymptotiquement : variance et erreurs de prévision sous-estimées, tests de Student invalidés
- pas de biais supplémentaire dans les estimations des coefficients
- les estimations peuvent ne pas être consistantes
- le TCL n'est plus valide (important si problème de normalité)

Une loi pour les résidus : la loi normale

Pour faire de l'inférence et des tests il faut une loi : loi normale

Non normalité des résidus :

- les estimations restent BLUE et n'affecte pas les résultats asymptotiques
- tests de significativité invalidés

Une loi pour les résidus : la loi normale

Pour faire de l'inférence et des tests il faut une loi : loi normale

Non normalité des résidus :

- les estimations restent BLUE et n'affecte pas les résultats asymptotiques
- tests de significativité invalidés


le TCL implique une loi asymptotiquement normale


TCL non valide en cas d'hétéroscédasticité

Sommaire

La multicolinéarité qu'est-ce que c'est ?

Multicolinéarité : problème qui survient lorsque certaines variables explicatives sont corrélées avec d'autres (ex : tendance commune)

Conséquences possibles :

- paramètres qui n'ont pas de sens avec des signes incohérents
- coefficients non significatifs (hausse variance) alors qu'une relation significative existe
- instabilité des coefficients en changeant d'échantillon ou en supprimant une variable
- résidus élevés

3 règles pour **soupçonner** la multicolinéarité

1. Si corrélation entre deux variables explicatives élevée (>0,8)
   Visualisation graphique sous R avec `corrplot::corrplot.mixed`
2. Règle de Klein : si carré de la corrélation entre deux variables explicatives est supérieure au R^2
3. Si un facteur d'inflation de la variance (FIV) est supérieur à 5 (ou 10 pour forte colinéarité)

   Dans le modèle $\forall t \in [\![1\,,\,T]\!]\,:\, y_t = \beta_0 + \sum_{n=1}^{N} \beta_n x_{n,t} + \varepsilon_t$ on a :

$$\mathbb{V}\left[\hat{\beta}_n\right] = \frac{\sigma^2}{T\mathbb{V}\left[x_n\right]} \underbrace{\frac{1}{1-R_n^2}}_{=FIV_n}$$

   avec R_n^2 le R^2 du modèle régressant x_n sur autres var. explicatives

   Sous R : `car::vif` sur modèle régression (objet `lm`)

Point aberrant : définitions

Définitions : les points aberrants

Un point aberrant (*outlier*) est une valeur extrême qui « s'éloigne » des autres observations. Dans la régression on distingue :

- les points à effet levier : valeurs extrêmes pour les variables explicatives
- les valeurs aberrantes : valeurs extrêmes pour la variable à prévoir

Point aberrant : définitions

Définitions : les points aberrants

Un point aberrant (*outlier*) est une valeur extrême qui « s'éloigne » des autres observations. Dans la régression on distingue :

- les points à effet levier : valeurs extrêmes pour les variables explicatives
- les valeurs aberrantes : valeurs extrêmes pour la variable à prévoir


Il est normal de trouver des valeurs extrêmes (ex loi normale : 1 obs sur 22 $>$ 2$\times$ l'écart type)


MCO minimise somme des carrés résidus $\Longrightarrow$ poids important si résidu élevé sur un *outlier* et peut altérer les estimations et conduire à des prévisions inexactes

Correction des *outliers*

3 méthodes usuelles pour corriger les *outliers* :

1. les supprimer de l'échantillon (difficile pour série temporelle)
2. chercher une variable économique qui les expliquent
3. rajouter des indicatrices $\rightarrow$ méthode privilégiée

Correction des *outliers*

3 méthodes usuelles pour corriger les *outliers* :

1. les supprimer de l'échantillon (difficile pour série temporelle)
2. chercher une variable économique qui les expliquent
3. rajouter des indicatrices $\rightarrow$ méthode privilégiée

 Un *outlier* peut être **inoffensif** ! $\rightarrow$ ajouter une indicatrice ne changera pas les estimations, l'ajustement (et le R^2) sera artificiellement amélioré mais pas la prévision !

Si *outlier* totalement aléatoire : pas de changement sur l'estimation

Si indicatrice corrélée à d'autres variables : ne pas la mettre va biaiser les coefficients

Détection *outliers*

Plusieurs méthodes et tests existent pour détecter les *outliers* :

- tests CUSUM (*cumulative sum*) ou MOSUM (*moving sum*) pour étudier la stabilité au cours du temps (changement de pente, etc.)
  Sous R : package `strucchange` fonction `efp` et `breakpoints`
- algorithmes automatiques : *Impulse Indicator Saturation* (IIS)
  Sous R : `isat` du package `gets`
- « À la main » : tester des indicatrices en ayant une connaissance sur les séries

Pas de soldes d'opinion, pas de prévision ?

Plusieurs recommandations :

- utiliser un modèle AR ou ARMA (valide pour séries stationnaires : théorème de Wold), **inconvénient** : ne détecte pas les retournements conjoncturels, erreurs de prévision pas toujours compréhensibles
- utiliser d'autres sources de données (pluviométriques, consommation électrique, etc.)
- prendre une prévision « par défaut » calculée par calage avec les autres modèles (ex. : prévision à un niveau désagrégé si l'on a la prévision au niveau agrégé et dans les autres niveaux désagrégés) ou en étudiant l'évolution tendancielle
- en « chaînant » les prévisions (ex. : prévision dans branche utilisée pour prévoir dans une autre branche)

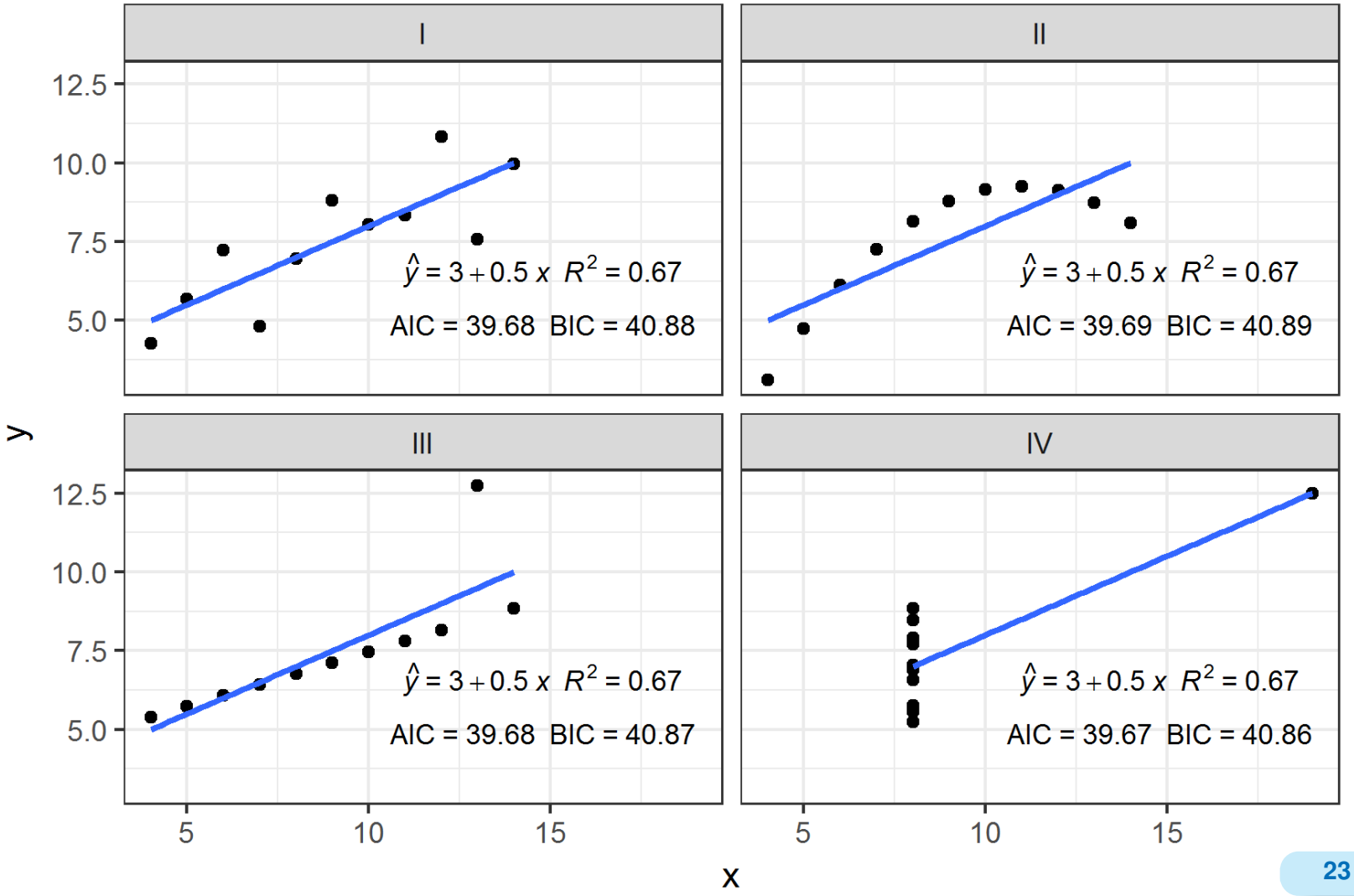
Conclusion

Pour construire un étalonnage il faut :

- des propriétés sur la variable à prévoir (stationnarité) et sur les variables explicatives (exogénéité) $\rightarrow$ bien comprendre comment sont construites les données pour choisir les variables les plus pertinentes
- choisir des variables qui ont du sens et faire attention à celles qui sont retenues (multicolinéarité)
- détecter et corriger les points atypiques qui ont un impact sur l'estimation (dépend des variables et du modèle utilisé)
- la relation doit être stable dans le temps
- vérifier des hypothèses sur les résidus (décorrélés, homoscédastiques et suivent une loi normale) : on peut parfois se passer de certaines hypothèses
- de préférence des signes « cohérents » et un nombre restreint de variables explicatives (attention au surajustement)

Conclusion

Toujours commencer par une analyse graphique : quartet d'Anscombe



Merci de votre attention !

Morgane Glotain morgane.glotain@insee.fr
Alain Quartier-la-Tente alain.quartier-la-tente@insee.fr