QUALITÉ D'UNE PRÉVISION (H6, J1, N6)
(10 / 06 / 2020, © Monfort, Dicostat2005, 2005-2020)

Certaines méthodes statistiques ont pour objectif d'apprécier la validité d'un ensemble de **prévisions**. La plupart de ces méthodes peuvent se décrire dans le cadre général définissant la notion de **prévision**.

On présente ici la notion de **qualité d'une prévision** dans le cadre du **modèle d'interdépendance**. En effet, diverses sciences (ou **domaines de connaissance**) utilisent souvent ce type de modèle, le plus souvent sous forme de **modèle dynamique** : physique (météorologie), écologie (évolutions migratoires), sociologie (démographie, économie).

(i) On suppose que la **forme structurelle** d'un modèle d'interdépendance est non linéaire (statique ou dynamique), ie que :

(1) $f(\eta, \xi) = \varepsilon_0$, avec $E\,\varepsilon_0 = 0$, $V\,\varepsilon_0 = \Sigma_0$,

où ξ est le vecteur des K **variables exogènes** (éventuellement retardées), η le vecteur des G **variables endogènes** (éventuellement retardées), ε_0 le vecteur des G perturbations aléatoires et $f : \mathbf{R}^K \times \mathbf{R} \mapsto \mathbf{R}^G$ une fonction vectorielle donnée.

(ii) Si le modèle (1) admet la **forme réduite** :

(2) $\eta = g(\xi) + \varepsilon$, avec $E\,\varepsilon = 0$, $V\,\varepsilon = \Sigma$,

et si g est estimée par $\tilde{g}$ à l'aide des **matrices d'observation** $X \in M_{NK}(\mathbf{R})$ de ξ et $Y \in M_{NG}(\mathbf{R})$ de η, on peut définir un **prédicteur conditionnel** à $\xi = x$ de η (ou, plus rigoureusement, un prédicteur de $E\,\eta = g(\xi)$) selon (cf **prédiction**) :

(3) $\tilde{\eta} = \tilde{g}(x)$.

(iii) Trois types de problèmes sont généralement étudiés :

(a) celui de l'**adéquation** de $\tilde{\eta}$ pr à η lorsque η est observée selon Y. Pour apprécier la qualité de $\tilde{\eta}$, on utilise une **distance entre va**, notée δ, ce qui définit eg une va D tq :

(4) $D^2 = ||\tilde{\eta} - \eta||^2$.

Si l'on pose $y_n\tilde{} = \tilde{g}(X_n')$, $\forall\, n \in N_N^*$, où X_n' désigne la n-ième observation de ξ (ie la n-ième ligne de X), la distance δ induit une distance entre $y_n\tilde{}$ et y_n , donc une distance D_N^2 entre $Y\tilde{}$ et Y, eg :

(5) $D_N^2 = N^{-1} \cdot \Sigma_{n=1}^{N} ||y_n\tilde{} - y_n||^2$.

C'est la « proximité » entre des statistiques tq D_N^2 et zéro qui est à l'origine des **tests d'adéquation** appropriés.

Ce type de problème est donc un **problème d'ajustement** (**estimation** et étude de sa qualité) appliqué à un modèle d'interdépendance (cf **qualité d'un ajustement**) : la prévision s'effectue pr à des valeurs déjà observées des variables ξ et η ;

(b) celui de l'**adéquation** de η^c pr à η lorsque η est observé selon une « valeur » Y_0 . Etant donné un « jeu » de M valeurs observées $X_0 \in M_{MK}$ (**R**) du vecteur ξ (ou **jeu d'observations**), on a :

(6) $\quad y_{0m}{}^{\sim} = g^{\sim}(X_{0m}), \quad \forall\, m \in N_M{}^*,$

où X_{0m} est la m-ième ligne de X_0 et où $g^{\sim}$ est toujours estimée à partir du **jeu de données** initial (X, Y).

On compare alors la prévision $Y_0{}^{\sim}$, faite en fonction de X_0 , avec la réalisation Y_0 effectivement observée. Une distance analogue à la précédente peut s'écrire :

(7) $\quad D_N^2 = N^{-1} . \Sigma_{n=1}^{N} ||y_{0m}{}^{\sim} - y_{0m}||^2.$

Cette notion de prévision, faite en fonction de X_0 , est la plus courante ;

(c) celui de la **comparaison** entre plusieurs prévisions d'un même vecteur η de variables endogènes, prévisions fondées ici sur divers prédicteurs :

$(c)_1$ prédicteurs calculés, à modèle donné, selon des méthodes différentes ;

$(c)_2$ prédicteurs calculés à partir de modèles différents (ie dont les spécifications diffèrent) ;

$(c)_3$ prédicteurs calculés à l'aide de divers **jeux de valeurs** X_0 , et ceci que l'on dispose des observations (X_0 , Y_0) correspondantes ou seulement des observations X_0 .

Le **choix de la meilleure prévision** s'effectue encore, le plus souvent, en comparant des **distances** du type précédent. On note que :

(a) les prévisions ainsi entendues sont des **prévisions « conditionnelles »** :

$(a)_1$ soit pr aux modèles (ie pr à leur **spécification**) ;

$(a)_2$ soit pr aux méthodes d'estimation ;

$(a)_3$ soit, plus couramment, pr aux (observations des) variables exogènes ;

(b) dans certains cas, la spécification (analytique ou statistique) d'un modèle dynamique permet d'améliorer la qualité des prévisions : ainsi, si les **perturbations** sont temporellement autocorrélées (cf **autocorrélation**), et si leur **matrice des covariances** est « correctement » estimée, on peut en tenir compte dans les formules définissant les prédicteurs ;

(c) pour procéder à un **test de qualité**, il existe divers **critères**, fondés sur des distances du type précédent (cf eg **coefficient de THEIL**, **méthode de NELSON**) ;

(d) une distance, tq celles décrites plus haut, s'associe souvent naturellement à une **fonctions de coût**, qui traduit l'inconvénient d'une prévision erronée. Ainsi, en notant (avec les mêmes notations que précédemment) :

(8) $u_n^{\sim} = y_n - y_n^{\sim}$ (resp $u_{0m}^{\sim} = y_{0m} - y_{0m}^{\sim}$)

l'erreur de prévision, cette **fonction de coût** (scalaire) $c : \mathbf{R}^G \mapsto \mathbf{R}$ est définie par :

(9) $e \mapsto c(e)$,

compte tenu de conditions usuelles tq :

(10)
$$c(0) = 0,$$
$$c(e_1) < c(e_2), \quad \forall (e_1, e_2) \text{ tq } ||e_1|| < ||e_2||.$$