Quelques outils statistiques pour la validation de modèles économiques

Valentin Patilea
Ensai

Journées IPR *Mathématiques et Sciences Sociales*

Plan de la présentation

1. Introduction – le problème économique
2. Le problème statistique – tester l'adéquation d'un modèle
3. Tests statistiques basés sur le lissage par noyau
4. Quelques remarques et conclusions

- Nécessité de modéliser l'économie: les économistes supposent que la satisfaction des *agents* est représentée par une fonction d'*utilité*.
- L'utilité de l'agent est fonction de la consommation et des heures travaillées: un arbitrage entre consommation et épargne et entre temps passé au travail et temps passé au loisir.
- L'utilité de l'agent dépends aussi des paramètres qui reflètent ses préférences.
- Les choix faits par l'agent sont obtenurs en maximisant la fonction d'utilité sous contrainte budgétaire et/ou monétaire.
- Les choix 'optimaux' faits par les ménages permettent de caractériser la demande en consommation.

- Dans nombre de cas, les variables intervenant dans l'utilité restent *incertaines* (gains d'une lotterie)
- Il 'semble' naturel de supposer qu'en présence d'incertitude l'agent maximise *l'espérance de son utilité future en tenant compte de l'information* dont il dispose au moment présent.
- Dans ce cas, la modélisation économique doit également tenir compte de la variabilité de l'utilité future autour de la moyenne conditionnelle à l'information disponible aujourd'hui.
- $\Rightarrow$ apparaît dès lors une notion de risque et d'aversion envers le risque.

- Le problème s'y pose tant pour les consommateurs que pour les producteurs qui cherchent à maximiser leurs profits sous contraintes (capacité de production,...) avec également une question d'incertitude.
- Pour les producteurs la fonction d'utilité est remplacée par la fonction de production.
- Les choix 'optimaux' des producteurs permettent de caractériser la partie offre des biens de consommation.

Un regard plus formel

- Optimisation sous contrainte
  - chercher les 'optima' comme solutions des *'conditions du* 1er *ordre'*
- L'information disponible par l'agent
  - modélisée par une *tribu*, plus précisement par la tribu des variables aléatoires observées par l'agent
- Espérance d'une fonction sachant une information
  - *Espérance mathématique conditionnelle à une tribu*
    - Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $\mathcal{G} \subset \mathcal{F}$. Soit X une variable aléatoire intégrable. Alors il existe une variable aléatoire Z, $\mathcal{G}-$mesurable et intégrable, telle que, $\forall U$ variable aléatoire bornée et $\mathcal{G}-$mesurable,
      $$\mathbb{E}[XU] \;=\; \mathbb{E}[ZU].$$
      On note alors $Z = \mathbb{E}[X|\mathcal{G}]$. La variable Z est unique presque sûrement (p.s.)

Un regard *encore* plus formel

- Une démarche de modélisation économique pourra conduire à affirmer
  - Il existent des vecteurs aléatoires $X \in \mathbb{R}^q$, $Y \in \mathbb{R}^d$ dont les réalisations sont observées par l'agent.
  - Il existe une fonction $g(X, Y; \theta)$ à valeurs réelles, avec $\theta \in \Theta \subset \mathbb{R}^p$.
  - La fonction g est connue et déduite d'une fonction d'utilité (ou production, ou...) et ses dérivées de 1er ordre.
  - Le vecteur θ contient des paramètres qui modélisent les caractéristiques des agents (aversion envers le risque,...).
  - La vraie valeur θ_0 du vecteur des paramètres est inconnue.
  - La valeur θ_0 est l'unique valeur de $\theta \in \Theta$ telle que

$$\mathbb{E}[g(Y, X; \theta) \mid \mathcal{G}] = 0, \quad p.s.$$

    où $\mathcal{G} = \sigma(X)$ (la tribu engendrée par le vecteur X).
- Question: comment vérifier si la modélisation économique est correcte?

- Les données: $n-$échantillon indép. $(\varepsilon_1, X_1), \cdots, (\varepsilon_n, X_n)$ de

$$(\varepsilon, X) \in \mathbb{R}^{1+q}$$

- En général, les variables $\varepsilon_1, \cdots, \varepsilon_n$ ne sont pas observées, elle sont à estimer à partir des données et avec l'aide du modèle.

- Avec les notations précedentes: $\varepsilon = g(Y, X; \theta_0)$.

- Le problème statistique: tester l'hypothèse statistique

$$\mathbb{E}\left[\varepsilon | X\right] = 0 \quad \text{p.s.}$$

- Ce problème est équivalent à un problème de test d'adéquation d'un modèle contre une hypothèse non paramétrique dans un ensemble très riche de modèles de régression.

Exemples

- Régression (non) linéaire

$$Y = m(X; \theta_0) + \varepsilon \quad \text{avec} \quad \mathbb{E}(\varepsilon \mid \mathrm{X}) = 0$$

  - La fonction $m(\cdot)$ est connue (exp. linéaire, logistique,...)
  - Le vecteur des paramètres θ_0 est inconnu est doit s'estimer à l'aide des données.

- Régression quantile

$$Y = q(X; \theta_0) + \eta$$

avec $\mathbb{P}(\eta \leq 0 \mid X) = \tau \in (0, 1)$,

$$\varepsilon = \mathbb{I}\{Y \leq q(X; \theta_0)\} - \tau$$

  - La fonction $q(\cdot)$ est connue.
  - Le vecteur des paramètres θ_0 est inconnu.

- Même si les réalisations de ε peuvent s'estimer à partir des données, la grande difficulté pour tester

$$\mathbb{E}(\varepsilon \mid X) = 0, \qquad \text{presque sûrement},$$

vient du fait qu'en général *la forme de l'espérance conditionnelle est inconnue.*

- IDEE 1: Soit $w(\cdot) > 0$ une fonction de poids.

Alors $\mathbb{E}(\varepsilon \mid X) = 0$ p.s. ssi

$$\mathbb{E}[\mathbb{E}^2(\varepsilon \mid X) w(X)] = 0,$$

et ceci est une condition de moment *marginal (ou non conditionnel)*.

- Un choix 'confortable': $w(X)$ est la densité de X.

- IDEE 1b: $\mathbb{E}(\varepsilon \mid X) = 0$ p.s. si et seulement si

$$\mathbb{E}[\varepsilon \mathbb{E}(\varepsilon \mid X) w(X)] = 0.$$

- IDEE 2: $\mathbb{E}(\varepsilon \mid X) = 0$ p.s. si et seulement si

$$\mathbb{E}[\varepsilon \mathbf{1}_{\{X \leq x\}}] = 0, \quad \forall x \in \mathbb{R}^q.$$

- IDEE 3: $\mathbb{E}(\varepsilon \mid X) = 0$ p.s. si et seulement si

$$\mathbb{E}[\varepsilon \exp(\langle x, X \rangle)] = 0, \quad \forall x \in D \subset \mathbb{R}^q.$$

Une formulation uniforme

Plus généralement on peut écrire

$$\mathbb{E}(\varepsilon \mid X) = 0 \quad \text{p.s.} \quad \Leftrightarrow \quad \mathbb{E}[\varepsilon\pi(X, x)] = 0, \quad \forall x \in D \subset \mathbb{R}^q.$$

avec

- IDEE 1b : $\pi(X, x) = \mathbb{E}(\varepsilon \mid X)w(X)$ une fonction inconnue function dépendant uniquement de X
- IDEE 2 : $\pi(X, x) = \mathbf{1}_{\{X \leq x\}}$, $x \in \mathbb{R}^q$
- IDEE 3 : $\pi(X, x) = \exp(\langle x, X \rangle)$, $x \in D \subset \mathbb{R}^q$.

Idée pour tester $\mathbb{E}(\varepsilon \mid X) = 0$ p.s.

- La caractérisation

$$\mathbb{E}(\varepsilon \mid X) = 0 \quad \text{p.s.} \quad \Leftrightarrow \quad \mathbb{E}[\varepsilon \pi(X, x)] = 0, \quad \forall x \in D \subset \mathbb{R}^q.$$

et loi des grands nombres nous suggère la solution pour la méthodologie de test.

  - Approcher $\mathbb{E}[\varepsilon \pi(X, x)]$ à l'aide d'une somme construite avec les observations:

$$M_n(x) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, \pi(X_i, x)$$

  - Vérifier si $M_n(x)$ est proche de zéro pour *tous* les $x \in D$.

- Par exemple, avec l'IDEE 3, on définit le processus stochastique

$$\widehat{M}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i \exp(\langle x, X_i \rangle), \qquad x \in D \subset \mathbb{R}^q.$$

- L'estimation $\widehat{\varepsilon}_i$ est obtenue dans le modèle, e.g.

$$\widehat{\varepsilon}_i = Y_i - m(X_i, \widehat{\theta}),$$

où $\widehat{\theta}$ est un estimateur $\sqrt{n}-$asymptotiquement normal de θ_0.

- *Il reste a définir ce que signifie $\widehat{M}_n(x)$ proche ou loin de zéro.*

- Avec un 'bon' estimateur $\widehat{s}^2(x)$ de la variance de $M_n(x)$ dans le main, on pourra montrer que

$$\widehat{W} = \sqrt{n}\frac{\widehat{M}_n}{\widehat{s}} \Rightarrow \mathbb{G}, \quad \text{si} \quad n \to \infty,$$

où $\mathbb{G}$ est un processus Gaussian centré.

- On peut utiliser des fonctionnelles du processus $\widehat{W}(x)$, $x \in D$, pour vérifier l'hypothèse de départ:
  - (statistique de test type Kolmogorov-Smirnov) $\sup_x |\widehat{W}(x)|$
  - (statistique de test type Cramér-von Mises) $\int_D \widehat{W}^2(x) d\mu(x)$,

- Sous l'hypothèse que $\mathbb{E}(\varepsilon \mid X) = 0$ p.s., les lois asymptotiques des statistiques de test KS et CvM sont données par les lois des fonctionnelles correspondantes du processus $\mathbb{G}$.
  On pourra ainsi dire ce que signifie $\widehat{M}_n(x)$ est *loin* de zéro.

- Des résultats similaires peuvent s'obtenir avec l'IDEE 2 et

$$M_n(x) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \, \mathbf{1}_{\{X_i \leq x\}},$$

$$\widehat{M}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i \mathbf{1}_{\{X_i \leq x\}}, \qquad x \in \mathbb{R}^q.$$

- Inconvenients de l'approche
  - Les valeurs critiques nécessaires pour décider si $\widehat{M}_n$ est proche ou loin de zéro sont difficile à calculer, ils demandent des procédures numériques.

- Rappel:

$$\mathbb{E}(\varepsilon \mid X) = 0 \;\; p.s. \Leftrightarrow \mathbb{E}[\varepsilon \mathbb{E}(\varepsilon \mid X) w(X)] = 0,$$

avec $w(\cdot)$ la densité de X.

- La loi de grands nombre suggère

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(X_i) \approx \mathbb{E}[\varepsilon \mathbb{E}(\varepsilon \mid X) w(X)]$$

ou $\phi(x) = \mathbb{E}(\varepsilon \mid X = x) w(x)$

- La fonction $\phi(\cdot)$ est inconnue et doit être estimée à partir des données.

Comment estimer $\mathbb{E}(\varepsilon \mid X)w(X)$?

- Soit $\phi(\cdot)$ une function mesurable continue en $x \in \mathbb{R}^q$.
- Soit $K(\cdot)$ une fonction noyau définie sur $\mathbb{R}^q$
  - bornée, intégrable d'intégrale 1 sur $\mathbb{R}^q$
  - $\|x\|^q K(x) \to 0$ si $\|x\| \to \infty$.
- Par exemple la densité $N(0,1)$, triangle ou quadratique sur $[-1,1]$,...
- Soit $h > 0$ un paramètre réel qui tend vers 0; h est appelé la *fenêtre*.
- Soit $K_h(x) = h^{-q}K(x/h)$.

- Alors, par le Lemme de Bochner, on a

$$\lim_{h \downarrow 0}(K_h \star \phi)(x) = \phi(x).$$

La convergence est uniforme en x si f est uniformement continue.

- Appliquons le Lemme de Bochner avec $\phi(x) = \mathbb{E}(\varepsilon \mid X = x)w(x)$.

- Alors

$$\begin{aligned}
(K_h \star \phi)(x) &= \int_{\mathbb{R}^q} K_h(x-u)\mathbb{E}(\varepsilon \mid X = u)w(u)du \\
&= \mathbb{E}[K_h(x-X)\mathbb{E}(\varepsilon \mid X)] \\
&= \mathbb{E}[K_h(x-X)\varepsilon] \\
&\to \mathbb{E}(\varepsilon \mid X = x)w(x), \quad \text{si} \quad h \downarrow 0.
\end{aligned}$$

- On peut donc prendre $h = h(n)$ tel que $h \downarrow 0$ lorsque $n \to \infty$ et utiliser l'approximation

$$\phi(X_i) \approx \sum_{j=1}^{n} \varepsilon_j K_h(X_i - X_j) \approx \sum_{j=1, j\neq i}^{n} \varepsilon_j K_h(X_i - X_j).$$

- On considère la statistique 'idéale'

$$T_n = nh^{q/2}\left[\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\varepsilon_i\,\varepsilon_j K_h\left(X_i - X_j\right)\right]$$

- Rappel: $\varepsilon_i = g(Y_i, X_i; \theta_0)$
- En réalité on devra se contenter de la statistique

$$\widehat{T}_n = nh^{q/2}\left[\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\widehat{\varepsilon}_i\,\widehat{\varepsilon}_j K_h\left(X_i - X_j\right)\right]$$

avec $\widehat{\varepsilon}_i = g(Y_i, X_i; \widehat{\theta})$.

- Eliminer la diagonale ($i = j$) permet d'avoir une statistique centrée lorsque l'hypothèse testée est vraie.

Comment ça fonctionne en pratique?

- Si l'hypothèse

$$\exists!\ \theta_0 \quad \text{tel que} \quad \mathbb{E}[g(Y, X; \theta_0) \mid X] = 0 \ \ p.s.$$

  est valide (c'est-à-dire, le modèle économique est correct), alors
  - les estimateurs usuels de θ_0 convergent vers θ_0 à la vitesse $n^{-1/2}$
  - $\widehat{\varepsilon}_i \approx \varepsilon_i$
  - la statistique $\widehat{T}_n$ aura la même loi asymptotique que T_n, c'est-à-dire $\widehat{T}_n$ sera *asymptotiquement gaussienne.*

- Si le modèle économique n'est pas correct, alors
  - les estimateurs usuels de θ_0 convergent vers un $\overline{\theta}$
  - $\widehat{\varepsilon}_i \approx g(Y, X; \overline{\theta})$ mais $\mathbb{E}[g(Y, X; \overline{\theta}) \mid X] \neq 0$
  - la statistique $\widehat{T}_n$ explosera vers $+\infty$

La procédure en pratique?

- On construit $\widehat{T}_n$
- On construit un estimateur $\widehat{V}_n^2$ de la variance de $\widehat{T}_n$
- La procédure de test: *"Rejeter H_0 lorsque $\widehat{T}_n/\widehat{V}_n \geq z_{1-\alpha}$"*
- Par exemple, avec un niveau 0,05, on rejete si $\widehat{T}_n/\widehat{V}_n \geq 1,645$.

Remarques sur la procédure de test

- Pour l'instant, les résultats théoriques ont une nature asymptotique. Avec 'un nombre insuffisant' des données, les seuils pour le rejet ou le non rejet peuvent être très imprécis.
  - Une solution: le bootstrap
- La procédure souffrira lorsque q, la dimension de X, est grande – le fléau de la dimension.
- La méthodologie peut se généraliser dans des situations 'non standard' (données manquantes, données fonctionnelles,...)

Une possible solution contre le fléau de la dimension

Lemma.

Soit $X \in \mathbb{R}^q$ *and* $\varepsilon \in \mathbb{R}$, $\mathbb{E}|\varepsilon| < \infty$.

A) Les affirmations suivantes sont équivalentes:

(i) $\forall \beta \in \mathbb{R}^q$ *non aléatoire avec* $\|\beta\| = 1$,

$$\mathbb{E}(\varepsilon \mid \langle X, \beta \rangle) = 0 \quad \textit{presque sûrement.}$$

(ii) $\mathbb{E}(\varepsilon \mid X) = 0$ *presque sûrement.*

B) Si X *est borné et* $\mathbb{P}\left[\mathbb{E}(\varepsilon \mid X) = 0\right] < 1$, *alors l'ensemble*

$$S = \{\beta \in \mathbb{R}^q : \|\beta\| = 1,\ \mathbb{E}(\varepsilon \mid \langle X, \beta \rangle) = 0 \textit{ presque sûrement}\}$$

est un ensemble rare sur la sphère unité $\{\beta \in \mathbb{R}^q : \|\beta\| = 1\}$.

En guise de conclusion

- Pour répondre à une question ‘simple’ venant de la modélisation économique, nous avons fait appel à une palette large d’outils mathématiques et informatiques
  - Analyse: mesure et intégration, transformée de Fourier, théorie de l’approximation des fonctions, espace de Hilbert,...
  - Algèbre: formes quadratiques, normes matricielles,...
  - Probabilités: loi de grands nombres, théorème central limit, processus stochastique,...
  - Informatique: Générateurs de nombres aléatoires, logiciels de calcul (Scilab, R, Matlab,...)
- Les problèmes économiques soulèvent très souvent des questions ‘simples’ qui posent des problèmes ‘compliqués’ aux mathématiques, en particulier à la statistique.