REGRESSION SUR UNE VARIABLE FONCTIONNELLE : ESTIMATION, TESTS DE STRUCTURES.

ALLAKERE HORMO MAXIME

January 23, 2021

INTRODUCTION

Les problèmes statistiques liés à l'étude de variable fonctionnelles connaissent depuis quelques années un intérêt grandissant dans la littérature, l'arrivée du Big Data et l'abondance des données ont motivés le développement et la recherche autours de ce thème, c'est par exemple des données météorologiques, médicales, imagerie satellite ne sont que quelques exemples illustrant le grand nombre et la diversité des données de nature fonctionnelles auxquelles le statistien est peut-être confronté. C'est une des raisons pour lesquelles un nouveau champ de la Statistique, dédié à l'étude de données fonctionnelles, a suscité un fort engouement au début des années quatre-vingt. Nous allons tout au long du document évoquer des termes mathématiques à l'exemple de régression, variable fonctionnelle etc. Le but de notre travail étant de vulgariser au maximum ce langage mathématique et faire comprendre le sens réel de ce langage abstrait pour le grand public, nous allons tout de suite commencer à définir quelques termes et expliquer le choix du sujet d'étude. Avant d'aborder la notion de variable fonctionnelle, une variable tout court est l'expression d'une valeur qui change en fonction d'un élément qui peut-être le temps par exemple.

Exemple : un serveur gagne 10€ l'heure et gagne en plus du salaire horaire un pourboire, le pourboire peut varier d'un client à un autre et donc la valeur du gain fait est par exemple : 10 + Pourboire, par définition notre pourboire est notre variable parce qu'il varie d'un client à un autre. L'exemple étant certes élémentaire, il était important de comprendre la notion de variable avant de définir les termes suivants : Une variable est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie (l'exemple d'un enregistrement sonore). L'observation d'une variable fonctionnelle est

appelée donnée fonctionnelle. La régression est une méthode d'analyse statistique permettant d'approcher une variable à partir d'autres qui lui sont corrélées, il existe plusieurs types de régressions entre autres : régression linéaire, régression linéaire multiple, régression non paramétrique, régression locale etc. Pour finir, un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique dite hypothèse nulle en fonction d'un jeu de données (échantillon).

PROBLEMATIQUE CONCRETE POUR VARIABLE FONCTIONNELLE

Reconnaissance vocale. — Dans le domaine de la linguistique, le problème de la reconnaissance vocale est un sujet d'actualité. L'objectif est de pouvoir retranscrire phonétiquement des mots et des phrases prononcés par un individu. Pour cela, il est auparavant nécessaire de procéder a une étape d'étalonnage qui consiste à recueillir différents enregistrements sonores pour chaque phonème. Les données obtenues sont intrinsèquement de nature fonctionnelle. Elles résultent de la discrétisation, sur une grille très fine, des signaux sonores enregistres au cours du temps. On n'étudie pas directement ces signaux mais plutôt les log-periodogrammes correspondants.La prise en compte de la nature fonctionnelle des données est un atout supplémentaire dans la comparaison de ces courbes puisqu'elle permet de dégager a la fois la forme générale du log-périodogramme et les variations d'amplitudes et de fréquences de ses oscillations plus fines.

Nous avons différents champs d'applications ou l'on peut-être confronté à des données de nature fonctionnelle, en voici quelques exemples :

- Dans le domaine de la médecine on peut observer l'utilisation de la statistique fonctionnelle au travers d'études de phonèmes différents tels que l'´évolution de certains cancers, l'effet placebo, l'activité cardiaque,les mouvements du genou pendant l'effort sous contrainte ou certaines déformations de la cornee.

- Depuis quelques années le secteur de la génétique est en plein essor. Grâce aux progrès effectués au niveau des appareils et des méthodes de mesure, les biologistes parviennent à faire plusieurs mesures de l'expression des gènes au cours du temps. Ces mesures ont pour objectif de permettre une meilleure compréhension de la fonction des gènes et des interactions entre certains de leurs effets (par exemple les phénomènes de régularisation d'une substance par une autre). On s'intéresse aussi à l'identification de groupes de gènes responsables de l'évolution d'un phénomène biologique complexe observé au cours du temps. Récemment, plusieurs méthodes de statistique

fonctionnelle ont été appliquées aux profils d'expression des gènes au cours du temps.

- Dans le domaine de l'économétrie on est confronté à de nombreux phénomènes que l'on peut modéliser par des variables fonctionnelles. On peut citer par exemple des travaux étudiant la dynamique de l'index mensuel de production de denrées périssables, la prédiction de la consommation électrique, la volatilité de marchés financiers, le rendement d'une entreprise, l'évolution du prix d'un objet au cours d'enchères.

LES METHODES D'ANALYSE DE DONNEES FONCTIONNELLES

Il existe plusieurs méthodes d'analyse des données fonctionnelles telles que : l'analyse factorielle (méthode basée sur l'étude d'elements propres de différents opérateurs), analyse exploratoire d'un échantillon de variables fonctionnelles (analyse des paramètres de centralités tels que la moyenne, la médiane et le mode), la régression pour variables fonctionnelles, nous baserons notre étude sur cette dernière méthode.

REGRESSION POUR VARIABLES FONCTIONNELLES

L'étude de modèles de régression adaptés à des données fonctionnelles est un domaine important de la statistique fonctionnelle. On y retrouve des situations tres différentes suivant que la variable explicative, la variable réponse ou les deux variables sont de nature fonctionnelle. Le premier modèle de régression considéré est le modèle linéaire fonctionnel, introduit et étudie par Ramsay et Dalzell (1991) puis Hastie et Mallows (1993), dans lequel on suppose que l'opérateur de régression est linéaire puis des modèles fonctionnels plus généraux ont été proposés.

Cependant, la régression sur variable fonctionnelle peut aussi être réalisée à partir de l'estimation d'autres quantités liées à la distribution conditionnelle de Y sachant X. Lorsque la régression met en évidence un lien linéaire significatif entre les variables x et y, ce résultat est parfois interprété, à tort, en termes d'influence ou de causalité de la variable x sur la variable y. Une méthode statistique ne peut, à elle seule, établir un lien de causalité entre deux variables. La causalité entre deux variables est un lien complexe à mettre en évidence, qui demande, entre autres, un plan expérimental spécifique, la répétabilité des résultats dans le temps, ainsi que sur divers échantillons. La régression linéaire ne peut pas être employée dans toutes les situations. Pour être utilisée cette méthode nécessite que les données satisfassent trois critères :

- La relation entre les deux variables doit être globalement linéaire, au moins grossièrement. C'est pour cette raison, qu'il faut toujours représenter graphiquement les données avec une droite de régression avant de choisir la méthode d'analyse.

- Les réponses doivent être indépendantes. C'est le plan d'échantillonnage qui renseigne sur cette condition. Si les données proviennent d'individus ou d'unités expérimentales différentes, elles sont généralement indépendantes. En revanche, si la variable indépendante est temporelle, les données ne sont sans doute pas indépendantes. Par exemple, si les réponses correspondent à des taux de glycémie mesurés quotidiennement sur un même patient, alors les réponses ne sont pas indépendantes.

- Les résidus doivent suivre une loi statistique. Et donc pour vérifier de fois ces conditions, nous faisons appel aux tests statistiques.

Modèles de régression sur variable fonctionnelle

Un modèle de régression s'écrit donc de cette façon : $Y = r(X) + \epsilon$

la variable Y est la valeur réelle tandis que la variable explicative X est à valeur dans un espace semi-métrique (E,d) de dimension infinie. ϵ représente le résidu.

Modèles paramétriques et non-paramétriques de régression sur variable fonctionnelle

Nous allons présenter des modèles de régressions sur variable fonctionnelle au travers des estimateurs paramétriques et non-paramétriques.

Considérons les trois équations suivantes :

(1) $Y = aX + b + \epsilon, X(\mu, \gamma^2)$ et $N(0, \sigma^2)$
(2) $Y = aX + b + \epsilon, E(\epsilon|X) = 0$
(3) $Y = r(X) + \epsilon, r \in C(\mathbb{R})$ et $E(\epsilon|X) = 0$

L'équation (1) est une équation paramétrique, l'équation (3) est une équation non-paramétrique tandis que la (2) est une équation paramétrique et non paramétrique selon les deux points de vue suivant :

- Non-Paramétrique si la loi du couple (X,Y)est supposé appartenir à un espace indexé par un nombre fini de paramètres réels.

- Paramétrique si l'opérateur de régression r est supposé appartenir à un espace indexé par un nombre fini de paramétriques réels.

Test de structure

Nous avons souligné le fait qu'il existe très peu de résultats concernant les tests de structures en régression sur variable fonctionnelle, on ne trouve que très peu d'articles scientifiques qui développent le sujet.

Il existe pas une approche théorique générale permettant de tester si un modèle possède une structure particulière : linéaire, à indice simple etc. Nous allons donc nous pencher sur le cas le plus simple ou l'on teste une hypothèse nulle soit H0 et voir si cette hypothèse est acceptée ou rejetée, en cas de rejet l'hypothèse alternative H1 est acceptée.

Tester si $r = r_0$ - le cas ou l'on souhaite tester l'hypothèse nulle.

$H_0 : (\mathbb{P}(r(X) = r_0(X)) = 1)$

Ou r_0 est un operateur connu contre l'alternative locale H1 :

$H_1 : (||r - r_0||_{L^2(wdP_x)} \geq \eta_n)$

Cela revient à faire un test de non effet sur l'échantillon, la façon dont η_n tend vers 0 reflète la capacité de notre test à détecter les différences de plus en plus petites entre r et r_0 lorsque n croit.

Pour finir, la mise en œuvre d'un test statistique nécessite en dehors des hypothèses de trouver une valeur seuil au delà de laquelle on rejette l'hypothèse nulle, les termes de variance sont difficiles à estimer et de plus utiliser la loi asymptotique pour obtenir des quantiles donne souvent des mauvais résultats.