Politechnika Gdańska
Wydział Fizyki Technicznej i Matematyki Stosowanej

Anna Wieżel
Nr albumu: 132540

Funkcjonalne Modele Liniowe

Praca magisterska
na kierunku MATEMATYKA
w zakresie MATEMATYKA FINANSOWA

Praca wykonana pod kierunkiem
dra hab. Karola Dziedziula
Katedra Analizy Matematycznej i Numerycznej

Wrzesień 2015

Oświadczenie kierujcego prac

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje si do przedstawienia jej w postpowaniu o nadanie tytułu zawodowego.

Data Podpis kierujcego prac

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowizujcymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur zwizanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załczon wersj elektroniczn.

Data Podpis autora (autorów) pracy

Streszczenie

coś

Słowa kluczowe

funkcjonalna analiza danych, dane funkcjonalne, funkcjonalne modele liniowe, test istotności

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka
11.2 Statystyka

Klasyfikacja tematyczna

62 Statistics
62-07 Data analysis
62J12 Generalized linear models

Tytuł pracy w jzyku angielskim

Functional Linear Models

Spis treści

Wstęp

Odpowiednik testu istotności dla prostego modelu regresji = F-test (+ t-test) [patrz: artykuł]

Rozdział 1

Preliminaria

1.1. Dane funkcjonalne. Elementy losowe

Przestrzenią funkcyjną nazywać będziemy przestrzeń liniową funkcji z dowolnego zbioru X do zbioru Y.

Definicja 1.1.1 *[FV]*
Zmienną losową $\mathcal{X}$ nazywamy ***zmienną funkcjonalną*** *wtedy i tylko wtedy, gdy przyjmuje wartości w nieskończenie wymiarowej przestrzeni (przestrzeni funkcyjnej). Obserwację χ zmiennej $\mathcal{X}$ nazywamy* ***daną funkcjonalną*** *(ang. functional data).*

Jeśli zmienna funkcjonalna $\mathcal{X}$ (odpowiednio obserwacja χ) jest krzywą zachodzi $\mathcal{X} = \{\mathcal{X}(t),\ t \in T\}$ (odp. $\chi = \{\chi(t),\ t \in T\}$), gdzie $T \subset \mathbb{R}$. Taką zmienną funkcjonalną możemy zatem utożsamiać z procesem stochastycznym z nieskończenie wymiarową przestrzenią stanów.
W szczególności, zmienna funkcjonalna może być powierzchnią, czyli dwuwymiarowym wektorem krzywych - wtedy analogicznie T będzie dwuwymiarowym zbiorem indeksów tj. $T \subset \mathbb{R}^2$ - lub dowolnie wymiarowym wektorem krzywych.

W pracy rozważać będziemy zmienne funkcjonalne przyjmujące wartości w przestrzeni Hilberta.

Zakładamy, że zmienne X_n, Y_n są scentrowanymi zmiennymi losowymi przyjmującymi wartości w przestrzeni Hilberta L^2. Oznaczając przez X (analogicznie Y) losową funkcję o tym samym rozkładzie co X_n (Y_n) wprowadzamy operatory

$$C(x) = \mathbb{E}[\langle X, x\rangle\, X], \quad \Gamma(x) = \mathbb{E}[\langle Y, x\rangle\, Y], \quad \Delta(x) = \mathbb{E}[\langle X, x\rangle\, Y].$$

...

1.2. Funkcjonalne modele liniowe

Rozróżniamy 3 postaci funkcjonalnych modeli liniowych

(i) pełen model funkcjonalny (ang. *the fully functional model*)

$$Y(t) = \int \beta(s,t)X(s)ds + \varepsilon(t),$$

(ii) model z odpowiedzią skalarną (ang. *the scalar response model*)

$$Y = \int \beta(s)X(s)ds + \varepsilon,$$

(iii) model z odpowiedzią funkcyjną (ang. *the functional response model*)

$$Y(t) = \beta(t)x + \varepsilon(t).$$

Rozważmy pełen model funkcjonalny postaci

$$\mathbf{Y}(t) = \int \beta(s,t)\mathbf{X}(s)ds + \boldsymbol{\varepsilon}(t),$$

gdzie

$$\mathbf{Y}(t) = \begin{bmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_N(t) \end{bmatrix}, \mathbf{X}(s) = \begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_N(s) \end{bmatrix}, \boldsymbol{\varepsilon}(t) = \begin{bmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \varepsilon_N(t) \end{bmatrix}.$$

Rozdział 2

Test istotności w funkcjonalnym modelu liniowym

2.1. Idea ogólna

Jednym z podstawowych testów na efektywność modelu jest test istotności zmiennych objaśniających. Zarówno jak i w przypadku klasycznego modelu liniowego w przypadku funkcjonalnego modelu liniowego badamy zerowanie się funkcji β, tj.

$$\mathrm{H}_0: \quad \boldsymbol{\beta} = 0 \quad \text{przeciw} \quad \mathrm{H}_A: \quad \boldsymbol{\beta} \neq 0.$$

Zauważmy, że przyjęcie H_0 nie oznacza braku związku między zmienną objaśnianą a objaśniającą. Prowadzi jedynie do stwierdzenia braku zależności liniowej.

Zakładamy, że zmienna objaśniana Y_n, zmienne objaśniające X_n i błędy ε_n są scentrowanymi zmiennymi losowymi przyjmującymi wartości w przestrzeni Hilberta L^2. Oznaczając przez X (analogicznie Y) losową funkcję o tym samym rozkładzie co X_n (Y_n) wprowadzamy operatory

$$C(x) = \mathbb{E}[\langle X, x\rangle\, X], \quad \Gamma(x) = \mathbb{E}[\langle Y, x\rangle\, Y], \quad \Delta(x) = \mathbb{E}[\langle X, x\rangle\, Y].$$

Przez $\hat{C}$, $\hat{\Gamma}$, $\hat{\Delta}$ oznaczamy ich estymatory, np.

$$\hat{C}(x) = \frac{1}{N}\sum_{n=1}^{N} \langle X_n, x\rangle\, X_n.$$

Definiujemy również wartości i wektory własne C i Γ

$$C(\upsilon_k) = \lambda_k \upsilon_k, \quad \Gamma(u_j) = \gamma_j u_j,$$

których estymatory będziemy oznaczać $(\hat{\lambda}_k, \hat{\upsilon}_k)$, $(\hat{\gamma}_j, \hat{u}_j)$.
Test obejmuje obcięcie powyższych operatorów na podprzestrzenie skończenie wymiarowe. Podprzestrzeń $\mathcal{V}_p = \mathrm{span}\{\upsilon_1, ..., \upsilon_p\}$ zawiera najlepsze przybliżenia X_n, które są liniowymi kombinacjami pierwszych p głównych składowych (FPC). Metodą głównych składowych wyznaczamy p największych wartości własnych operatora $\hat{C}$ tak, że $\hat{\mathcal{V}}_p = \mathrm{span}\{\hat{\upsilon}_1, ..., \hat{\upsilon}_p\}$ zawiera najlepsze przybliżenie X_n. Analogicznie $\mathcal{U}_q = \mathrm{span}\{u_1, ..., u_q\}$ zawiera przybliżenia $\mathrm{span}\{Y_1, ..., Y_N\}$.
Z równości

$$Y(t) = \int \beta(s,t)X(s)ds + \boldsymbol{\varepsilon}(t)$$

wynika $\Delta = \beta C$ i dla $k \leqslant p$ mamy

$$\beta(\upsilon_k) = \lambda_k^{-1}\Delta(\upsilon_k).$$

Stąd, β zeruje się na $\text{span}\{\upsilon_1, ..., \upsilon_p\}$ wtedy i tylko wtedy, gdy $\Delta(\upsilon_k) = 0$ dla każdego $k = 1, ..., p$. Zauważmy, że

$$\Delta(\upsilon_k) \approx \hat{\Delta}(\upsilon_k) = \frac{1}{N}\sum_{n=1}^{N} \langle X_n, \upsilon_k\rangle Y_n.$$

Skoro zatem $\text{span}\{Y_1, ..., Y_N\}$ są dobrze aproksymowane przez $\mathcal{U}_q$, to możemy ograniczyć się do sprawdzania czy

$$\left\langle \hat{\Delta}(\upsilon_k), u_j \right\rangle = 0, \quad k = 1, ..., p, \quad j = 1, ..., q. \tag{2.1}$$

Jeśli H_0 jest prawdziwa, to dla każdego $x \in \mathcal{V}_p$, $\beta(x)$ nie należy do $\mathcal{U}_q$. Co znaczy, że żadna funkcja Y_n nie może być opisana jako liniowa kombinacja X_n, $n = 1, ..., N$. Statystyka testowa powinna zatem sumować kwadraty iloczynów skalarnych (2.1). Poniższe twierdzenia prowadzą do wyznaczenia statystyki

$$\hat{T}_N(p, q) = N\sum_{k=1}^{p}\sum_{j=1}^{q} \hat{\lambda}_k^{-1}\hat{\gamma}_j^{-1}\left\langle \hat{\Delta}(\hat{\upsilon}_k), \hat{u}_j \right\rangle^2, \tag{2.2}$$

która zbiega według rozkładu do rozkładu χ^2 z pq stopniami swobody.
Przy czym

$$\left\langle \hat{\Delta}(\hat{\upsilon}_k), \hat{u}_j \right\rangle = \left\langle \frac{1}{N}\sum_{n=1}^{N} \langle X_n, \hat{\upsilon}_k\rangle Y_n, \hat{u}_j \right\rangle = \frac{1}{N}\sum_{n=1}^{N} \langle X_n, \hat{\upsilon}_k\rangle \langle Y_n, \hat{u}_j\rangle$$

oraz $\lambda_k = \mathbb{E}\langle X, \upsilon_k\rangle^2$ i $\gamma_j = \mathbb{E}\langle Y, u_j\rangle^2$.

Uwaga 2.1.1 *Oczywistym jest, że jeśli odrzucamy H_0, to $\beta(\upsilon_k) \neq 0$ dla pewnego $k \geqslant 1$. Jednak ograniczając się do p największych wartości własnych, test jest skuteczny tylko jeśli β nie zanika na którymś wektorze υ_k, $k = 1, ..., p$. Aczkolwiek takie ograniczenie jest intuicyjnie niegroźne, ponieważ test ma za zadanie sprawdzić czy główne źródła zmienności Y mogą być opisane przez główne źródła zmienności zmiennych X.*

Schemat przebiegu testu

1. Sprawdzamy założenie o liniowości metodą *FPC score predictor-response plots.*
2. Wybieramy liczbę głównych składowych p i q metodami *scree test* oraz *CPV.*
3. Wyliczamy wartość statystyki $\hat{T}_N(p, q)$ (2.2).
4. Jeśli $\hat{T}_N(p, q) > \chi^2_{pq}(1 - \alpha)$, to odrzucamy hipotezę zerową o braku liniowej zależności. W przeciwnym razie nie mamy podstaw do odrzucenia H_0.

...

2.2. Formalizm

Założenia:

1. Trójka $(Y_n, X_n, \varepsilon_n)$ tworzy ciąg niezależnych zmiennych losowych o jednakowym rozkładzie, takich że ε_n jest niezależne od X_n oraz

$$\mathbb{E}X_n = 0, \quad \mathbb{E}\varepsilon_n = 0,$$

$$\mathbb{E}\|X_n\|^4 < \infty \quad \text{i} \quad \mathbb{E}\|\varepsilon_n\|^4 < \infty.$$

2. Wartości własne operatorów C oraz Γ spełniają, dla pewnych $p > 0$ i $q > 0$

$$\lambda_1 > \lambda_2 > ... > \lambda_p > \lambda_{p+1}, \quad \gamma_1 > \gamma_2 > ... > \gamma_q > \gamma_{q+1}.$$

Twierdzenie 2.2.1 *Jeśli spełnione są H_0 i powyższe Założenia, to $\hat{T}_N(p,q) \xrightarrow{d} \chi^2_{pq}$ przy $N \to \infty$.*

Twierdzenie 2.2.2 *Przy powyższych Założeniach oraz jeśli $\langle \beta(\upsilon_k), u_j \rangle \neq 0$ dla pewnych $k \leqslant p$ oraz $j \leqslant q$, to $\hat{T}_N(p,q) \xrightarrow{P} \chi^2_{pq}$ przy $N \to \infty$.*

Dowody...

Rozdział 3

Przykład zastosowania

Magnetometer data...

Dodatek A

Kod w R

...

Bibliografia

[B] D. Bosq, *Linear Processes in Function Spaces*. Springer, 2000.

[FV] F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis. Theory and practice*. Springer, 2006.

[HK] Lajos Horváth, Piotr Kokoszka, *Interference for Functional Data with Applications*. Springer, 2012.

[I] INTERMAGNET `http://www.intermagnet.org/index-eng.php`

[K08] P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, *Testing for lack of dependence in the functional linear model*. Canadian Journal of Statistics, 2008, 36, 207-222.

[RS05] J. O. Ramsay, B. W. Silverman, *Functional Data Analysis*. Springer, 2005.