Analiza Danych Funkcjonalnych — projekt końcowy

Jakub Zbrzezny

3 czerwca 2021

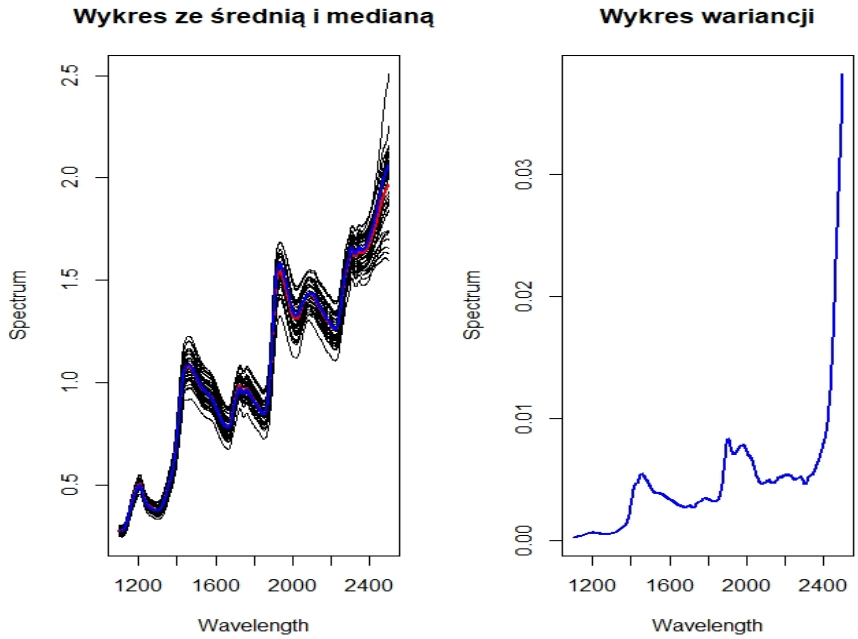
1 Wstęp

Moim zadaniem jest na podstawie zbioru danych **Biscuit** z pakietu **fds**, przeanalizowanie danych funkcjonalnych metodami statystyki opisowej, następnie stworzenie modeli regresji i ich porównanie, a później utworzenie modelu klasyfikacji na podstawie danych funkcjonalnych dla grup $\{fat < 18\}$ i $\{fat \geqslant 18\}$.

Dane pochodzą z pewnego eksperymentu, który polegał na zmianie składu kawałków ciasta biszkoptowego. Badano dwa takie zbiory, pierwszym z nich jest zbiór kalibracyjny, a drugim z nich jest zbiór predykcyjny. Były tworzone oraz rozpatrywane jako dwa odrębne zbiory, przy różnych okazjach, i nie są wynikiem przypadkowego (ani żadnego innego) podziału większego zbioru.
Zbiorami kalibracyjnymi są zbiory **nirc**, **labc** (czyli wszystkie o nazwach kończących się na literę "c"), natomiast predykcyjnymi są **nirp**, **labp** (czyli wszystkie o nazwach kończących się na literę "p").
W zbiorach **nirp**, **nirc**, mamy informacje na temat spektrów dotyczących ciastek.
Natomiast, w zbiorach **labp**, **labc**, można znaleźć informacje o składzie ciastek, takie jak tłuszcz, sacharoza, mąka, woda.

2 Analiza danych funkcjonalnych metodami statystyki opisowej

Przedstawię teraz wykres danych funkcjonalnych wraz z średnią oraz medianą oraz wykres wariancji dla zbioru **nirp**.



Na wykresie ze średnią i medianą, czerwona linia przedstawia średnią, natomiast niebieska, medianę. Widzimy, że średnia prawie wszędzie pokrywa się z medianą. Średnia znajduje się mniej więcej pośrodku między wszystkimi obserwacjami. Ponadto, nie widzimy żadnych obserwacji odstających.
Natomiast, w przypadku wykresu wariancji, mamy dwa wyraźne skoki wariancji, a potem od długości fali spektrometru równej 2400, wariancja rośnie wykładniczo.

3 Modele regresji i ich porównanie

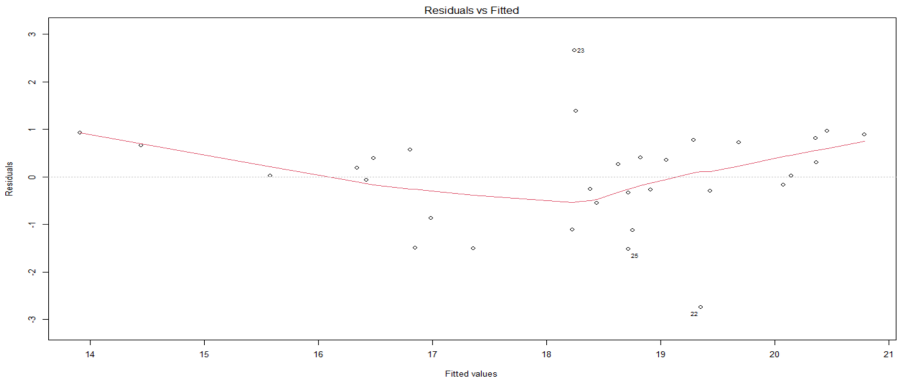
Teraz stworzę różne modele regresji z skalarną zmienną objaśnianą oraz funkcjonalną zmienną objaśniającą ze zbioru **nirp**. Dla każdej z 4 skalarnych zmiennych ze zbioru **labp**: **tłuszcz**, **sacharoza**, **mąka**, **woda**, stworzę 2 modele z jedną konkretną zmienną objaśnianą w zależności od funkcjonalnej zmiennej objaśniającej: **model z reprezentacją bazową**, z funkcji **fregre.lm** oraz **model z nieparametryczną estymacją jądrową**, z funkcji **fregre.np**.
Porównam jakości dopasowania modeli oraz zbiór współczynników istotnych statystycznie dla różnych zmiennych objaśnianych.

3.1 Tłuszcz

Dla modelu z reprezentacją bazową, otrzymałem następujące wyniki:

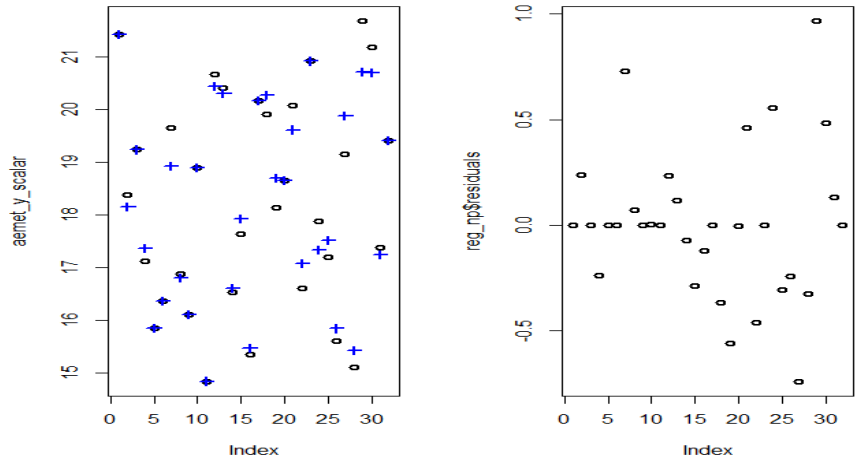
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      18.257812   0.199312  91.604  < 2e-16 ***
aemet_x1.bspl4.1  1.704676   0.569081   2.995 0.005950 **
aemet_x1.bspl4.2 -2.069009   0.477552  -4.333 0.000196 ***
aemet_x1.bspl4.3  1.489363   0.573204   2.598 0.015228 *
aemet_x1.bspl4.4 -0.240747   0.350949  -0.686 0.498792
aemet_x1.bspl4.5 -0.005584   0.113857  -0.049 0.961261
```

Zatem 3 pierwsze współczynniki są istotne, natomiast pozostałe nie są istotne statystycznie. Otrzymany współczynnik R^2, wynosi 0.7361.
Zbadam dopasowanie modelu za pomocą wykresu rezyduów względem dopasowanych wartości:



Widzimy, że krzywa najlepszego dopasowania wyraźnie odbiega wokół zera oraz widać, że w środku, więcej rezyduów jest poniżej zera niż powyżej oraz na prawej stronie wykresu, przeważają rezydua powyżej zera, zatem model nie jest dobrze dopasowany. Obserwacje nr 22, 23, 25, są obserwacjami odstającymi.

Natomiast w przypadku modelu z nieparametryczną estymacją jądrową, współczynnik R^2 jest równy 0.9681478. Sprawdzę dopasowanie modelu za pomocą wykresu rezyduów:



Widać, że rezydua są równomiernie rozłożone wokół zera, zatem model jest dobrze dopasowany.

3.2 Sacharoza

Dla modelu z reprezentacją bazową, tutaj już wszystkie współczynniki są istotne statystycznie, ale współczynnik R^2 wyszedł mniejszy: 0.7231.
Natomiast w przypadku modelu z nieparametryczną estymacją jądrową, współczynnik R^2 jest trochę większy niż dla tłuszczu: 0.9737446.

3.3 Mąka

Dla modelu z reprezentacją bazową, tutaj znów wszystkie współczynniki są istotne statystycznie, a współczynnik R^2 wyszedł równy: 0.7737.
Natomiast w przypadku modelu z nieparametryczną estymacją jądrową, współczynnik R^2 wynosi: 0.9828765.

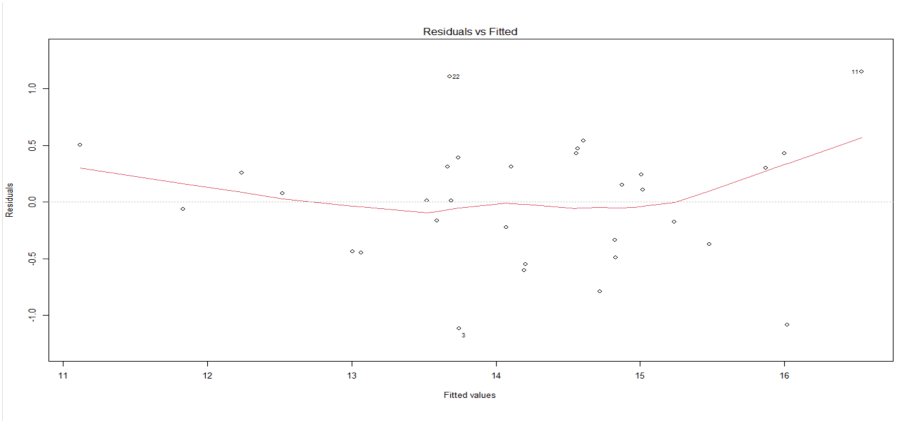
3.4 Woda

Dla modelu z reprezentacją bazową, tylko pierwszy współczynnik nie jest istotny statystycznie, a pozostałe są, a współczynnik R^2 wyszedł wyraźnie większy niż dla wcześniejszych zmiennych: 0.8402.
Natomiast w przypadku modelu z nieparametryczną estymacją jądrową, współczynnik R^2 wyszedł również największy spośród wszystkich zmiennych ze zbioru **labp**: 0.9874809.

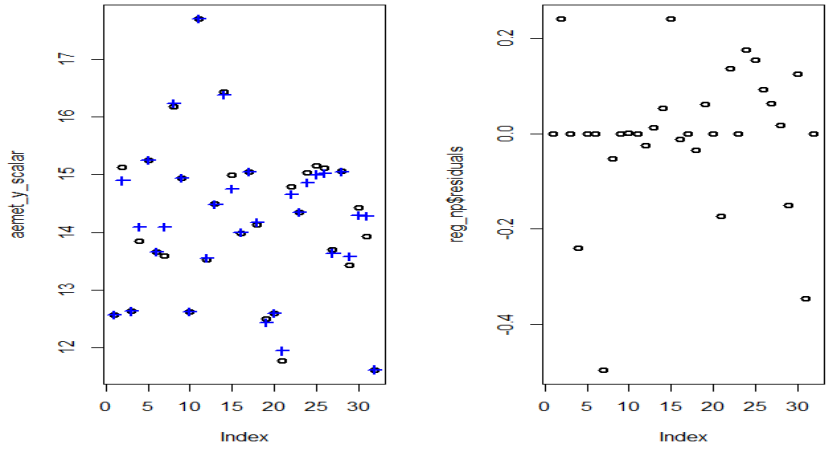
3.5 Modele regresji - podsumowanie

Podsumowując, najlepiej sprawdził się model przewidujący zawartość wody w zależności od spektrów dotyczących ciastek, ponieważ dla obu różnych modeli, współczynnik dopasowania R^2 wyszedł największy. Ogólnie, zawsze lepiej był dopasowany model z nieparametryczną estymacją jądrową niż model z reprezentacją bazową.
Co więcej, zbiory istotnych statystycznie współczynników były różne. Najmniej było w przypadku **tłuszczu**: 3 z 5, a najwięcej było w przypadku **sacharozy** i **wody**: wszystkie.
Przedstawię dopasowania tych najlepszych modeli:

Model z reprezentacją bazową:



Model z nieparametryczną estymacją jądrową:



Widzimy, że w przypadku 1-go modelu, jedynie na krańcach przedziału widzimy odchyłki od zera, natomiast na środku wykres wygląda dobrze. Natomiast, jeśli chodzi o 2-gi model, to generalnie rezydua są rozłożone symetrycznie wokół zera, jedynie 2 rezydua odbiegają od symetrii.

4 Model klasyfikacji na podstawie danych funkcjonalnych dla grup $\{fat < 18\}$ i $\{fat \geqslant 18\}$

Następnie, stworzyłem model klasyfikacji na podstawie danych funkcjonalnych dla grup $\{fat < 18\}$ i $\{fat \geqslant 18\}$. Żeby zadziałała funkcja klasyfikująca dla danych funkcjonalnych **classif.kernel**, dokonałem przekształcenia zbioru **labp** na dane funkcjonalne poleceniem: **fdata(t(labp))**. Następnie zbiór podzieliłem na dwie rozpatrywane grupy tłuszczu. Zbiorem treningowym były dane o $32/2 = 16$ wierszach (ta połowa wszystkich wierszy została wybrano losowo funkcją **sample**). Natomiast zbiorem testowym były dane z pozostałych 16 wierszy. Później dokonałem klasyfikacji funkcją **classif.kernel**, a potem dokonałem predykcji na zbiorze testowym. Wyniki były następujące:

```
            predykcja
grupa_test 1 2
         1 2 4
         2 3 7
```

Prawdopodobieństwo, że wartość przewidywana była równa prawdziwej, wynosiło 56.25%, zatem było dość małe.