**METODY EKSPLORACJI DANYCH**

Laboratorium. Modele logitowe. Regresja logistyczna.

Prowadzący: Wykonali:

dr inż. Romuald Hoffmann pchor. Michał ADAMCZEWSKI

pchor. Mikołaj ADAMSKI

pchor. Przemysław SUJECKI

Zadanie 3.

Przez cały okres eksploatacji pewnego systemu operacyjnego (OS) zbierano dane dotyczące liczby błędów krytycznych wykrytych w tym czasie w oprogramowaniu. Zebrane obserwacje w układzie miesięcznym przedstawiono w tabeli 3.

Obraz zawierający tekst, numer, Równolegle, linia

Opis wygenerowany automatycznie

W zadaniu proszę:

1. Wyznaczyć zależność sumarycznej liczby błędów w okresie eksploatacji oprogramowania badanego systemu operacyjnego.

2. Na podstawie opracowanego modelu i przeprowadzonych obliczeń sformułować własne wnioski.

3. Wyniki analizy proszę zawrzeć w postaci sprawozdania, do którego proszę dodać jako załączniki wszystkie pliki z obliczeniami (obliczenia można przeprowadzić w dowolnie wybranym narzędziu.

Funkcja logistyczna jest nieliniową funkcją jednej zmiennej, którą z reguły jest zmienna 𝑡 odpowiadająca czasowi. Funkcja logistyczna, której wykresem jest krzywa przypominająca kształtem literę „S”, należy do rodziny krzywych sigmoidalnych (ang. sigmoid curve).

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Przyjmijmy do dalszych rozważań, że zapiszemy jako:

Obraz zawierający tekst, Czcionka, linia, biały

Opis wygenerowany automatycznie

Jak łatwo zauważyć, funkcja ta nie jest liniowa względem parametrów. Ideą metody Hotellinga jest takie przekształcenie funkcji logistycznej, aby uzyskać zależność liniową, w której występowałyby pewne proste funkcje parametrów.

Różniczkując funkcję logistyczną względem zmiennej objaśniającej otrzymujemy:

Obraz zawierający Czcionka, linia, pismo odręczne, tekst

Opis wygenerowany automatycznie

Stąd otrzymujemy:



Pamiętając o tym, że przyrost zmiennej objaśniającej t (np. w funkcji trendu) może być przyjmowany jako równy jedności mamy 𝑑𝑦 = Δ𝑦 (to wynika z założenia, że różniczka funkcji jest równa jej przyrostowi). Wówczas mamy:



Oznaczamy przez: przez , stąd

Wówczas ostatnie zależność ta przyjmuje postać:



Parametry 𝛽0 i 𝛽1 w powyższej zależności możemy oszacować metodą najmniejszych kwadratów, ponieważ jest to zależność liniowa względem parametrów. W ten sposób oszacujemy parametr 𝛽0 = 𝑎 logistycznego modelu oraz parametr 𝛽1 = , jako funkcję parametru 𝑘. Oznaczmy przez 𝛽̂0 i 𝛽̂1 oceny

parametrów 𝛽 i 𝛽1. Wówczas mamy:



W celu oszacowania parametru 𝑏 przekształcimy wzór:

Obraz zawierający Czcionka, linia, numer, tekst

Opis wygenerowany automatycznie

w następujący sposób:



Z powyższego mamy:



Całkiem naturalne wydaje się wiec przyjęcie za ocenę parametru ln(𝑏) takiej wartości, która dla oszacowanych wartości parametrów 𝑎 i 𝑘 (tzn. 𝑎̂ oraz 𝑘̂) minimalizuje wyrażenie:

Obraz zawierający Czcionka, linia, biały, diagram

Opis wygenerowany automatycznie

Otrzymujemy wówczas następujący estymator ln(𝑏):

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Jeżeli chcielibyśmy otrzymane metodą Hotellinga wzory na estymatory parametrów 𝑎, 𝑏 oraz 𝑘 zapisać za pomocą danych, to przyjmą one postać jak poniżej.

Biorąc pod uwagę, że mamy oraz zbiór obserwacji {𝑦𝑡∶ 𝑡 = 1, … , 𝑛} to mamy:

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Parametr a szacujemy ze wzoru:

Obraz zawierający Czcionka, linia, tekst, diagram

Opis wygenerowany automatycznie

Oszacowanie parametru 𝑘 ma postać:

Obraz zawierający Czcionka, tekst, linia, pismo odręczne

Opis wygenerowany automatycznie

Parametr b szacujemy ze wzoru:

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Modele logitowe stosujemy głownie w modelowaniu zmiennych jakościowych, inaczej zwanych dychotomicznych, które opisują fakt wystąpienia lub niewystąpienia analizowanego zjawiska; np. zakupu określonej rzeczy, korzystania/rezygnacji z określonej usługi, itd.

Wartość oczekiwana zmiennej objaśnianej może być interpretowana jako warunkowe prawdopodobieństwo realizacji danego zdarzenia przy ustalonych wartościach zmiennej objaśniającej lub zmiennych objaśniających (może być ich więcej). Wartość 𝑦̂𝑡 uważamy za oszacowanie prawdopodobieństwa wystąpienia intersującego zdarzenia. Wadą tego rozwiązania jest to, że praktycznie wartość 𝑦̂𝑡 może być spoza przedziału wartości [0, 1] i wobec tego użycie takiego modelu daje wątpliwy efekt. Aby temu zaradzić dokonuje się odpowiedniego przekształcenia prawdopodobieństwa z przedziału

[0, 1] w przedział (−∞, +∞) stosując tak zwaną transformację logitową postaci:

gdzie:

𝑝 – jest prawdopodobieństwem wystąpienia zdarzenia.

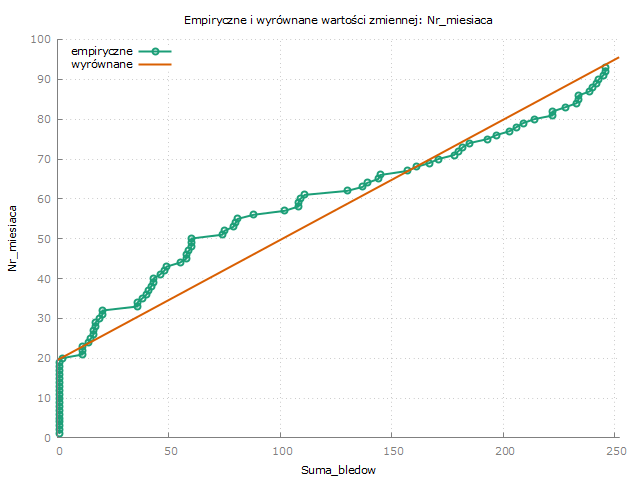
Iloraz nazywany jest szansą wystąpienia zdarzenia.

Prawdopodobieństwo 𝑝 szacujemy na podstawie zaobserwowanych danych dotyczących zmiennych 𝑌 i 𝑋. W praktyce empirycznej wartość prawdopodobieństwa 𝑝 szacujemy jako częstości empiryczne 𝑝𝑡 wystąpienia zdarzenia. Wobec tego dla każdego wystąpienia mamy transformację:

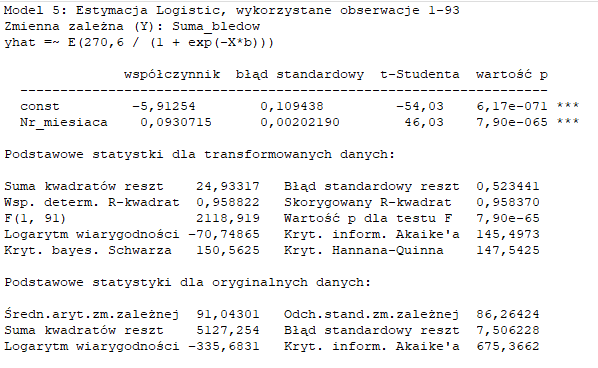
Stosując przekształcenie odwrotne do transformacji 𝐿 = mamy:

Prawdopodobieństwo 𝑝 szacujemy na podstawie zaobserwowanych danych dotyczących zmiennych 𝑌 i 𝑋. W praktyce empirycznej wartość prawdopodobieństwa 𝑝 szacujemy jako częstości empiryczne 𝑝𝑡 wystąpienia zdarzenia. Wobec tego dla każdego wystąpienia mamy transformację:

Wykres zależności sumarycznej liczby błędów na przestrzeni wszystkich miesięcy wygenerowany w GRETLU:



Wyniki:



Prognozy dla obserwacji od 54 do 93 (39 wartości).

Obraz zawierający tekst, diagram, Wykres, mapa

Opis wygenerowany automatycznie

Wyniki:

Obraz zawierający tekst, menu, zrzut ekranu, dokument

Opis wygenerowany automatycznie

**Wnioski:**

Z miesiąca na miesiąc liczby błędów w systemie przybywało co można zauważyć na pierwszym wykresie. Do 20. miesiąca sumaryczna liczba błędów wynosiła 2. Wraz z biegiem czasu w systemie pojawiały się coraz to większe ilości błędów. Największe ilości błędów zaczęły pojawiać się od 31. miesiąca, gdzie liczba błędów wynosiła 10. W 33. miesiącu pojawiła się największa liczba błędów na przestrzeni wszystkich obserwacji wynosząca 16. Może to oznaczać, że wraz z użytkowaniem systemu użytkownicy/testerzy zaczęli „wyłapywać” nowe błędy.

Sumaryczna liczba błędów wynosi 246. Obserwacje dla prognoz podzielone są następująco: 1-53 - obserwacje bazowe, 54-93 – prognozy.

Widzimy, że RMSE (pierwiastek błędu średniokwadratowego) dla prognoz wynosi 1,8236, więc można stwierdzić, że model przewiduje dobrze dalsze prognozy.

Podobnie sytuacja wygląda z MAPE (średni absolutny błąd procentowy), gdzie wynosi on 2,1049 %. Innymi słowy, prognozy różnią się średnio o 2% od rzeczywistych wartości.

Bibliografia:

[0] dr inż. Romuald Hoffmann, prof. WAT, Notatki dla studentów, Warszawa 2023.

[1] Larose D. T., Metody i modele eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa 2012.

[2] <https://scikit-learn.org/stable/>

[3] Fred Nwanganga, Mike Chapple, Praktyczne uczenie maszynowe w języku R, Wiley, Warszawa, 2022.

[4] Laurence Moroney, Sztuczna inteligencja i uczenie maszynowe dla programistów, Helion O’Reilly, 2021.

[5] <https://pl.wikipedia.org/wiki/Wikipedia:Strona_g%C5%82%C3%B3wna>

[6] Aurelie Geron, Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow, Helion O’Reilly, 2018