**METODY EKSPLORACJI DANYCH**

Laboratorium. Analiza regresji - regresja liniowa

Przekształcanie funkcji nieliniowych w równoważne liniowe

Budowa modelu na podstawie danych (ogólnie dostępnych)

Prowadzący: Wykonali:

dr inż. Romuald Hoffmannpchor. Michał ADAMCZEWSKI

pchor. Mikołaj ADAMSKI

pchor. Przemysław SUJECKI

## Zadanie nr 2.

## Badanie zależności między wagą, a obwodem brzucha (abdomen).

Wczytujemy plik .CSV za pomocą :



Rys. 1 Wczytanie pliku csv

Używając biblioteki:



Rys. 2 Biblioteka tidyverse - RStudio

Możemy teraz przejrzeć dane :



Rys. 3 Wydruk danych

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

Rys. 4 Zobrazowanie danych za pomocą funkcji glimpse

Podsumowanie danych otrzymujemy poprzez:



Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

Rys. 5 Podsumowanie wszystkich rekordów z pliku csv

Widać, że średnia i mediana zmiennej odpowiedzi są bardzo zbliżone co sugeruje normalny rozkład danych. Możemy to sprawdzić za pomocą histogramu:

Obraz zawierający tekst, linia, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

Rys. 6 Tworzenie histogramu

Obraz zawierający diagram, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznie

Rys. 7 Histogram Pct.BF

Widać jednak, że jakieś wartości są równe 0 wartość procentowa tkanki tłuszczowej nie może być równa 0 więc należy usunąć tą wartość traktując jako błąd.

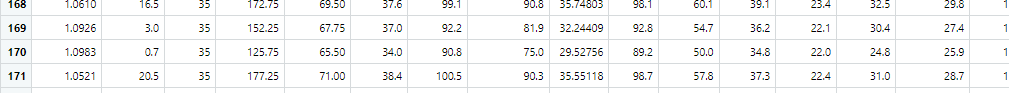
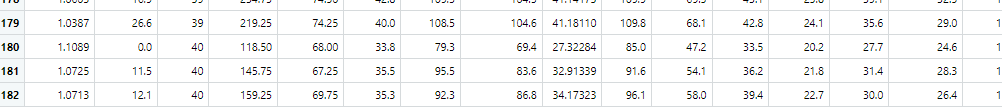


Rys. 8 Ograniczenie Pct.BF



Rys. 9 Rekordy, gdzie Pct.BF jest bliski zerze

Pokazuje nam, że to rekord 180 oraz 170.

Rys. 10 Rekordy z niskim Pct.BF



Rys. 11 Tworzenie histogramu z ograniczeniem Pct.BF

Obraz zawierający diagram, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznie

Rys. 12 Histogram z ograniczeniem Pct.BF

Możemy teraz stworzyć histogramy zmiennych predykatorów:

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Rys. 13 Tworzenie histogramu zmiennych predykatorów

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rys. 14 Histogramy zmiennych predykatorów

Można zobaczyć że np. age, czyli wiek, ma jednostajny rozkład, czyli reprezentuje pacjentów w różnym wieku.

Możemy teraz sprawdzić zależności pomiędzy zmiennymi za pomocą współczynnika korelacji Pearsona :



Rys. 15 Przypisanie zmiennej Facik\_correlations korelacji między zmiennymi

Wizualizujemy to za pomocą:

****

Rys. 16 Tworzenie wykresu korelacji między zmiennymi

Otrzymujemy wykres:

Obraz zawierający tekst, zrzut ekranu, Wielobarwność, design

Opis wygenerowany automatycznie

Rys. 17 Wykres korelacji między zmiennymi z pliku csv

Na podstawie tego wykresu możemy stwierdzić, że zmienna odpowiedzi jest najsilniej skorelowana z brzuchem, talią, wagą, biodrem oraz udem.

W takim razie zaczniemy budować model na podstawie naszych obserwacji:



Rys. 18 Tworzenie modelu ze zmiennymi Waist, Weight, Abdomen oraz Hip



Rys. 19 Podsumowanie modelu

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rys. 20 Wydruk wyników modelu

Jak widać dostaliśmy NA wszędzie na zmiennej Abdomen (Brzuch), jest to spowodowane tym że zachodzi silna korelacja między zmienną Abdomen oraz Waist (Talia), która wynosi 1,0. Należy więc usunąć jedną z nich (od nas zależy którą). Zdecydujemy się na usunięcie talii - czyli Waist.

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Rys. 21 Ponowne podsumowanie modelu

W takim razie teraz wyniki naszego modelu wyglądają tak:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rys. 22 Wydruk wyników bez zmiennej Waist

Pokazuje nam, że zmienna hip nie ma większego wpływu na nasz model więc zdecydujemy się na jej usunięcie.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rys. 23 Podsumowanie wyników bez zmiennej Hip

**Min** – o tyle model nie doszacował wartości w najgorszym wypadku.

**Max** – o tyle model przeszacował wartość w najgorszym wypadku.

**Mediana** – pokazuje, że generalnie przewiduje wartości które są odrobine mniejsze niż rzeczywiste.

Współczynniki:

Dla wagi oznacza, że w miarę zwiększania wagi o jednostkę, przewidywany procent tkanki tłuszczowej spada o około 0.13599 punktu procentowego. P-value < 0.05, co oznacza, że współczynnik ten jest statystycznie istotny.

Dla brzucha oznacza, że w miarę zwiększania obwodu talii o jednostkę, przewidywany procent tkanki tłuszczowej rośnie o około 0.97390 punktu procentowego. P-value < 0.05, co oznacza, że współczynnik ten jest statystycznie istotny.

**Residual standard error** - Średni błąd predykcji modelu wynosi około 4.385 punktu procentowego na 245 stopniach swobody. To miara, która wskazuje, jak bardzo przewidywane wartości różnią się od rzeczywistych wartości. U nas możemy odczytać, że różni się o 4 punkty procentowe.

**Multiple R-squared -** Około 72% zmienności w procentach tkanki tłuszczowej ("Pct.BF") jest wyjaśniane przez nasz model.

Podsumowując, model wydaje się być stosunkowo dobry, z wysokim wielokrotnym R-kwadratem, co sugeruje, że duży odsetek zmienności w procentach tkanki tłuszczowej jest wyjaśniany przez zmienne w modelu. Statystyka F potwierdza, że przynajmniej jedna zmienna jest istotna, a reszty modelu również są statystycznie istotne.

Wykonamy teraz kilka testów.

Pierwszym z nich będzie sprawdzenie zerowej średniej reszt:





Rys. 24 Obliczenie zerowej średniej reszt

Średnia jest bliska 0, więc test uznany jest za zaliczony.

Drugi test to rozkład normalny reszt:



Rys. 25 Tworzenie histogramu rozkładu normalnego reszt

Obraz zawierający diagram, Wykres, linia, zrzut ekranu

Opis wygenerowany automatycznie

Rys. 26 Histogram rozkładu normalnego reszt

Kolejnym testem którym się zajmiemy to test autokorelacji reszt:



Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Rys. 27 Obliczenie autokorelacji reszt

Wynik testu : 1.781588. Wartości między 1.5 a 2.5 sugerują, że prawdopodobnie brak jest autokorelacji natomiast wartości p-value są stosowane do oceny istotności statystyki testowej. P-value wynosi 0.064. Jeśli przyjmiemy poziom istotności na poziomie 0.05, to p-value > 0.05 sugeruje, że nie ma autokorelacji reszt.

Teraz uruchomimy test sprawdzający punktów wpływowych w danych poprzez funkcje odległości Cooka.



Rys. 28 Tworzenie wykresu funkcji odległości Cooka

Obraz zawierający tekst, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznie

Rys. 29 Wykres funkcji odległości Cooka

Widać kilka punktów wyróżniających się czyli punktów wpływowych. Takimi obserwacjami są np. obserwacja nr 40 albo 246 czy 32.









Rys. 30 Wyróżniające się dane

Porównując :

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rys. 31 Wydruk podsumowanych danych z pliku csv

## Badanie zależności między klatką piersiową, a talią.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rys. 32 Wydruk wyników ze zmiennymi Chest oraz Waist

Współczynniki:

Dla klatki piersiowej wynik oznacza, że w miarę zwiększania klatki piersiowej o jednostkę, przewidywany procent tkanki tłuszczowej spada o około 0.28981 punktu procentowego. P-value < 0.05, co oznacza, że współczynnik ten jest statystycznie istotny.

Dla talii wynik oznacza, że w miarę zwiększania obwodu talii o jednostkę, przewidywany procent tkanki tłuszczowej rośnie o około 2.23510 punktu procentowego. P-value < 0.05, co oznacza, że współczynnik ten jest statystycznie istotny.

**Residual standard error** - Średni błąd predykcji modelu wynosi około 4.619 punktu procentowego na 247 stopniach swobody. To miara, która wskazuje, jak bardzo przewidywane wartości różnią się od rzeczywistych wartości. U nas możemy odczytać, że różni się o 5 punktów procentowych.

**Multiple R-squared -** Około 70% zmienności w procentach tkanki tłuszczowej ("Pct.BF") jest wyjaśniane przez nasz model.

Podsumowując, model wydaje się być stosunkowo dobry, z wysokim wielokrotnym R-kwadratem, co sugeruje, że duży odsetek zmienności w procentach tkanki tłuszczowej jest wyjaśniany przez zmienne w modelu. Statystyka F potwierdza, że przynajmniej jedna zmienna jest istotna, a reszty modelu również są statystycznie istotne.

Wykonamy ponownie te same testy.

Zerowa średnia reszta:



Rys. 33 Obliczenie zerowej średniej reszt

Średnia ponownie jest bliska 0, więc test uznany jest za zaliczony.

Rozkład normalny reszt:

Obraz zawierający diagram, Wykres, linia, zrzut ekranu

Opis wygenerowany automatycznie

Rys. 34 Histogram rozkładu normalnego reszt

Test autokorelacji reszt:

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Rys. 35 Obliczenie autokorelacji reszt

Wynik testu : 1.84696. P-value wynosi 0.184.

Punkty wpływowe w danych:

Obraz zawierający tekst, Wykres, diagram, linia

Opis wygenerowany automatycznie

Rys. 36 Wykres funkcji odległości Cooka

Ponownie widać kilka punktów wyróżniających się czyli punktów wpływowych. Takimi obserwacjami są np. obserwacja nr 40, 126, 138 i 214.









Rys. 37 Wyróżniające się dane

Porównując:

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Rys. 38 Wydruk podsumowanych danych z pliku csv

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