WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

WYDZIAŁ CYBERNETYKI

SPRAWOZDANIE
Metody Eksploracji Danych

Temat laboratorium: **ANALIZA REGRESJI – DOBÓR ZMIENNYCH DO MODELU WIELOCZYNNIKOWEGO**

INFORMATYKA

(kierunek studiów)

INŻYNIERIA SYSTEMÓW – ANALIZA DANYCH

(specjalność)

Zespół:

Michał ŚLĘZAK
Szymon OLEŚKIEWICZ

Prowadzący laboratorium:

Dr inż. Romuald Hoffmann, prof. WAT

Warszawa 2025

Spis treści

Rozdział I. Zadanie 2 - Bodyfat

Zadanie

Zebrano pomiary 250 mężczyzn w różnym wieku (źródło: http://www.byu.edu/chhp). Zebrane dane zawierają poprawnie zmierzony % tłuszczu „Pct.BF" (%bodyfat) oraz inne zweryfikowane wartości zmiennych chrakteryzujących sylwetkę. Tab. 1 zawiera tylko fragment danych i stanowi tylko ich ilustrację. Pełny zestaw danych znajduje się w pliku o nazwie „MED-lab-1-Zadanie-2-Dobor zmiennych-bodyfat-dane-i-opis.txt"

Tabela 1. Fragment danych z pliku „MED-Lab-1-Zad-Dobor zmiennych-bodyfat-dane-i-opis.txt"

Gęstość	Pct.BF	Wiek	Waga	Wzrost	Szyja	Klatka	Brzuch	Talia
Density	Pct.BF	Age	Weight	Height	Neck	Chest	Abdomen	Waist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	33.543307
1.0853	6.1	22	173.25	72.25	38.5	93.6	83	32.677165
1.0414	25.3	22	154	66.25	34	95.8	87.9	34.606299
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	34.015748
1.034	28.7	24	184.25	71.25	34.4	97.3	100	39.370079
1.0502	20.9	24	210.25	74.75	39	104.5	94.4	37.165354
1.0549	19.2	26	181	69.75	36.4	105.1	90.7	35.708661
1.0704	12.4	25	176	72.5	37.8	99.6	88.5	34.842520
1.09	4.1	25	191	74	38.1	100.9	82.5	32.480315
1.0722	11.7	23	198.25	73.5	42.1	99.6	88.6	34.881890
1.083	7.1	26	186.25	74.5	38.5	101.5	83.6	32.913386
1.0812	7.8	27	216	76	39.4	103.6	90.9	35.787402
1.0513	20.8	32	180.5	69.5	38.4	102	91.6	36.062992
1.0505	21.2	30	205.25	71.25	39.4	104.1	101.8	40.078740
1.0484	22.1	35	187.75	69.5	40.5	101.3	96.4	37.952756
1.0512	20.9	35	162.75	66	36.4	99.1	92.8	36.535433
1.0333	29	34	195.75	71	38.9	101.9	96.4	37.952756
1.0468	22.9	32	209.25	71	42.1	107.6	97.5	38.385827
1.0622	16	28	183.75	67.75	38	106.8	89.6	35.275591

Proszę:

1) zbudować model pozwalający przewidzieć %bodyfat na podstawie innych zmiennych. Procent tłuszczu ciała każdego (%bodyfat, PctBF) znajduje się w drugiej kolumnie danych.
2) przed budową modelu proszę zaproponować/wybrać procedurę eliminacji zmiennych wraz z uzasadnieniem.
3) dodatkowo, na podstawie danych zaproponować model/-e dla innej/-ych zmiennej/-ych objaśnianej/-ących.
4) przygotować w formie pisemnej wyczerpujące sprawozdanie z wykonania zadania (!), które wraz z innymi plikami (wymaganymi do weryfikacji rozwiązania zadania) należy wstępnie umieścić w prywatnym notesie zespołu MS Teams w sekcji „Sprawozdania" w odpowiednio nazwanej stronie, np. Sprawozdanie z lab.1

I.1. Wykorzystane narzędzia i zależności

I.1.1. Modele

W celu wyznaczenia modeli liniowych wraz z ich parametrami na podstawie wykładów oraz wiedzy własnej przygotowano program w języku Python wykorzystujący biblioteki Pandas, NumPy, stastmodel, scikit-learn oraz scipy, które również zostały wykorzystane do obliczeń oraz wizualizacji wyników. Modele oraz ich parametry zostały wyznaczone za pomocą wyżej wspomnianych narzędzi, które oparte są o metodę MNK (Metoda Najmniejszych Kwadratów minimalizująca sumę kwadratów reszt RSS), gdzie wraz z dopasowaniem danych, wyznaczono wiele parametrów i metryk dla danego stworzonego modelu. Model liniowy jednej zmiennej wyraża się poniższym wzorem.

$$\hat{y} = a_1 x + a_0 \tag{1}$$

Należy wyznaczyć takie wartości parametrów $a_1 \; i \; a_0$, aby najlepiej dopasować przebieg prostej regresji do wartości danych. Aby tego dokonać, należy rozwiązać poniższe zadanie optymalizacyjne.

$$M(a_0, a_1) = \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2 \tag{2}$$

Przyrównując pochodne funkcji sumy kwadratów reszt po zmiennych $a_0 \; i \; a_1$ do zera, otrzymamy układ równań, który w postaci macierzowej będzie mieć następującą postać.

$$\begin{bmatrix} \sum_{i=1}^{n} {x_i}^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \tag{3}$$

Po lewostronnym przemnożeniu obu stron równania przez macierz odwrotną do stojącej przy macierzy parametrów, otrzymamy wzór na wartość macierzy tych parametrów.

$$\begin{bmatrix} \sum_{i=1}^{n} {x_i}^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} = \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} \tag{4}$$

Powyższy tok rozumowania można uogólnić dla regresji wielu zmiennych. W tym celu można wykorzystać zapis macierzowy, gdzie po przekształceniach otrzymamy ostatecznie wzór na wartości współczynników dla dowolnej liczby parametrów.

$$A = \begin{bmatrix} a_0 \\ \dots \\ a_k \end{bmatrix} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y \tag{5}$$

W powyższym wzorze macierze X oraz Y zostały wyznaczone na podstawie zbioru danych, takich jak poniżej.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \dots \\ y_k \\ \dots \\ y_n \end{bmatrix} \tag{6}$$

To właśnie tym przypadkiem będziemy się zajmować w ramach tego laboratorium – modelem liniowym z wieloma zmiennymi.

I.1.2. Metody doboru zmiennych

Pierwszą metodą doboru zmiennych, którą wybrano jest **metoda eliminacji wstecznej**. **Zdecydowano się na jej wybór ze względu na jej prostotę**. Zaczynamy budowę modelu od wszystkich zmiennych objaśniających. Następnie, usuwamy z listy cech zmienną, która daje najlepszą wartość miary oceny. Proces ten kontynuujemy aż do osiągnięcia zadanego kryterium. Zdecydowano o wykorzystaniu miary w postaci statystyki |t|. W każdym kroku usuwamy zmienną o najniższej wartości statystyki t-studenta |t|, która jest mniejsza od wartości krytycznej tej statystyki wziętej z tablic, jeśli jej wartość jest mniejsza od krytycznej wartości tej statystyki. Proces ten powtarza się aż wszystkie pozostałe współczynniki będą istotne statystycznie, czyli będą miały wartość statystyki |t| większą lub równą wartości krytycznej, którą odczytamy z tablic na podstawie poziomu istotności alfa oraz stopni swobody.

Drugą metodą doboru zmiennych, którą wykorzystano w zadaniu jest **metoda regularyzacji Lasso**. Polega ona na dodaniu kary do różnych parametrów modelu w celu zmniejszenia nadmiernego dopasowania. W regularyzacji modelu liniowego kara nakładana jest na współczynniki, które mnożą każdy z predyktorów. Poprzez nakładanie kar, współczynniki zerują się. Metoda Lasso rozszerza klasyczną regresję liniową o karę L1 na bezwzględne wartości współczynników. Lambda/Alfa to parametr regularyzacji kontrolujący siłę kary. Wraz ze wzrostem siły kary, Lasso „sprowadza” współczynniki do zera, dzięki temu eliminowane są nieistotne zmienne.

Należy jednak pamiętać, że zmienne „wejściowe” trzeba ustandaryzować, ponieważ kara działa na bezwzględne wartości współczynników. Bez standaryzacji zmienne o dużych skalach otrzymywałyby nieproporcjonalnie dużą karę w porównaniu do zmiennych o mniejszych skalach.

Budowa tego modelu polega na rozszerzeniu metody opartej na najmniejszych kwadratach o karę. Czyli do modelu zwykłej regresji liniowej musimy dodać poniższą część.

$$\alpha \cdot \sum |a_i|$$

Regularyzację Lasso można wyrazić także za pomocą poniższego wzoru, który, tak jak wspomniano, jest rozszerzeniem Metody Najmniejszych Kwadratów.

$$min(RSS) = \sum (Y_i - \hat{Y}_i)^2 + \alpha \cdot \sum |a_i|$$

Do budowy tego modelu używamy LassoCV, które samo dobiera nam optymalne alfa poprzez, w naszym wypadku, 5 krotną cross walidację oraz ocenę na podstawie maksymalizacji R^2.

Trzecią metodą doboru zmiennych, którą wykorzystano w zadaniu jest metoda **Forward Selection**. Polega na stopniowym dodawaniu zmiennych objaśniających do modelu, zaczynając od modelu bez predyktorów. Na każdym kroku wybierana jest zmienna, która maksymalizuje skorygowany współczynnik determinacji modelu. Proces ten zatrzymuje się, gdy dodanie kolejnej zmiennej nie poprawia już skorygowanego R^2, czyli różnica między skorygowanym współczynnikiem determinacji modelu z nową zmienną a tym skorygowanym współczynnikiem z poprzedniego modelu bez zmiennej jest mniejsza lub równa 0.

I.1.3. Ocena predykcji modelu

W celu oceny predykcji modelu wykorzystamy metodę post ante.
Predykcja post ante ma na celu zweryfikowanie poprawności przewidywanych przez model wartości do wartości rzeczywistych danych. Istnieje wiele metryk błędu prognozy post ante, poniżej znajduje się omówienie niektórych z nich.

Średni błąd predykcji ME pokazuje czy model systematycznie przeszacowuje (dodatni ME) lub nie doszacowuje (ujemny ME) wartości rzeczywistych danych. Wartości powinny być jak najbardziej bliskie zera. Opisuje go poniższy wzór.

$$ME = \frac{1}{m} \sum_{\tau=1}^{m} (y_\tau - y_\tau^p)$$

Średni błąd kwadratowy predykcji MSE pokazuje średnią kwadratową różnicę między wartościami rzeczywistymi a estymowanymi. Opisuje go poniższy wzór.

$$MSE = \frac{1}{m}\sum_{\tau=1}^{m}(y_\tau - y_\tau^p)^2$$

Średni błąd absolutny MAE mierzy przeciętną wielkość błędu bez uwzględniania jego kierunku, jego wartość jest zawsze dodatnia. Pokazuje jak bardzo przewidywane wartości są średnio odległe od danych rzeczywistych. Określony jest on poniższym wzorem.

$$MAE = \frac{1}{m}\sum_{\tau=1}^{m}|y_\tau - y_\tau^p|$$

Pierwiastek średniego błędu kwadratowego RMSE mocniej wyróżnia większe błędy przez podniesienie do kwadratu. Opisuje go poniższy wzór.

$$RMSE = \sqrt{\frac{1}{m}\sum_{\tau=1}^{m}(y_\tau - y_\tau^p)}$$

Średni procentowy błąd absolutny MAPE pokazuje jak duże są błędy w odniesieniu do rzeczywistych wartości, wyrażone w procentach. Opisuje go poniższy wzór.

$$MAPE = \frac{1}{m}\sum_{\tau=1}^{m}\left|\frac{y_\tau - y_\tau^p}{y_\tau}\right| \cdot 100\%$$

Jak już wcześniej wspomniano, do obliczeń wykorzystano narzędzia w postaci języka programowania Python, który, m.in. oblicza niektóre metryki.

Odchylenie standardowe modelu S_e uzyskamy pierwiastkując wariancję S_e^2.

$$S_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$$

Na podstawie wartości odchylenia standardowego S_e modelu jednej zmiennej można wyznaczyć odchylenia parametrów $a_0\ i\ a_1$, tak jak poniżej.

$$S_{a_0} = S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}$$

$$S_{a_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}$$

W celu zbadania jak dobrze wyznaczona prosta regresji oddaje przebieg zmiennej można wykorzystać współczynnik determinacji, który obliczany jest na podstawie poniższego wzoru. Współczynnik ten przyjmuje wartości od 0 do 1. Im bliżej 1, tym nasz model jest bardziej dopasowany.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

W celu zbadania jak dobrze wyznaczona prosta regresji oddaje przebieg zmiennej w zależności od liczby predyktorów można wykorzystać skorygowany współczynnik determinacji, który obliczany jest na podstawie poniższego wzoru. Pokazuje on procent zmiennej zależnej wyjaśniony przez model z uwzględnieniem kary za liczbę predyktorów p. Współczynnik ten przyjmuje wartości od 0 do 1. Im bliżej 1, tym nasz model jest bardziej dopasowany.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2) \cdot (n - 1)}{n - p - 1}$$

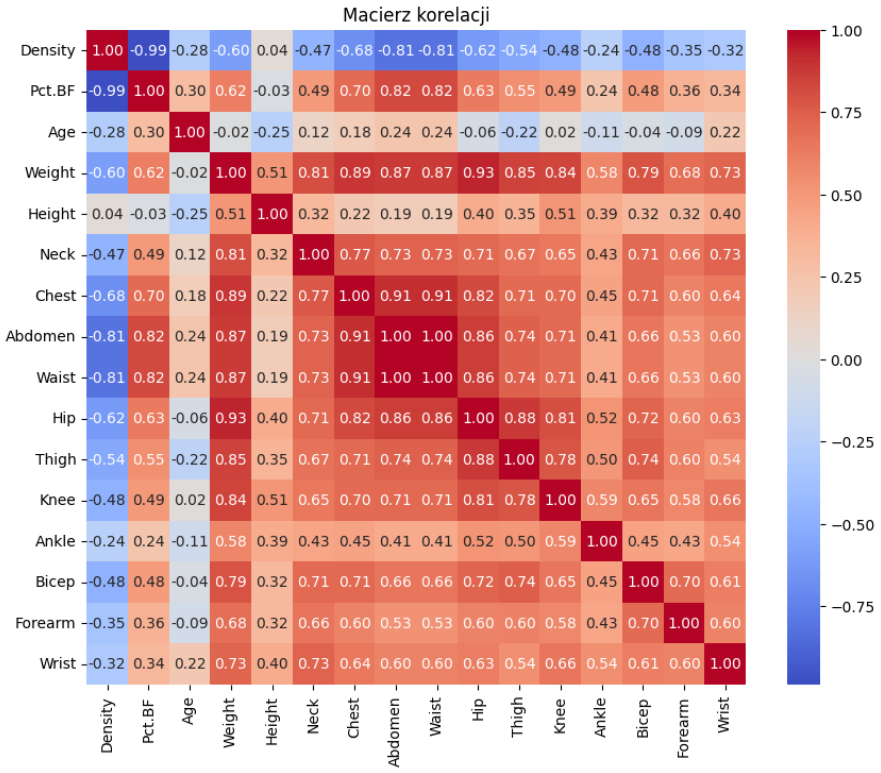
Następną miarą oceny dopasowania modelu do danych rzeczywistych jest współczynnik zmienności losowej, którego wzór zamieszczono poniżej. Współczynnik ten porównuje się z wartością krytyczną $W_e^* \in [0\%, 25\%]$, która gdy $W_e \leq W_e^*$ to model jest dobrze dopasowany.

$$W_e = \frac{S_e}{\bar{y}} \cdot 100\%$$

I.2. Eksploracja danych

I.2.1. Sprawdzenie korelacji między zmiennymi

W celu sprawdzenia korelacji między zmiennymi, wykorzystano macierz korelacji opartą o metodę Pearsona oraz stworzono jej wizualizację na mapie cieplnej.

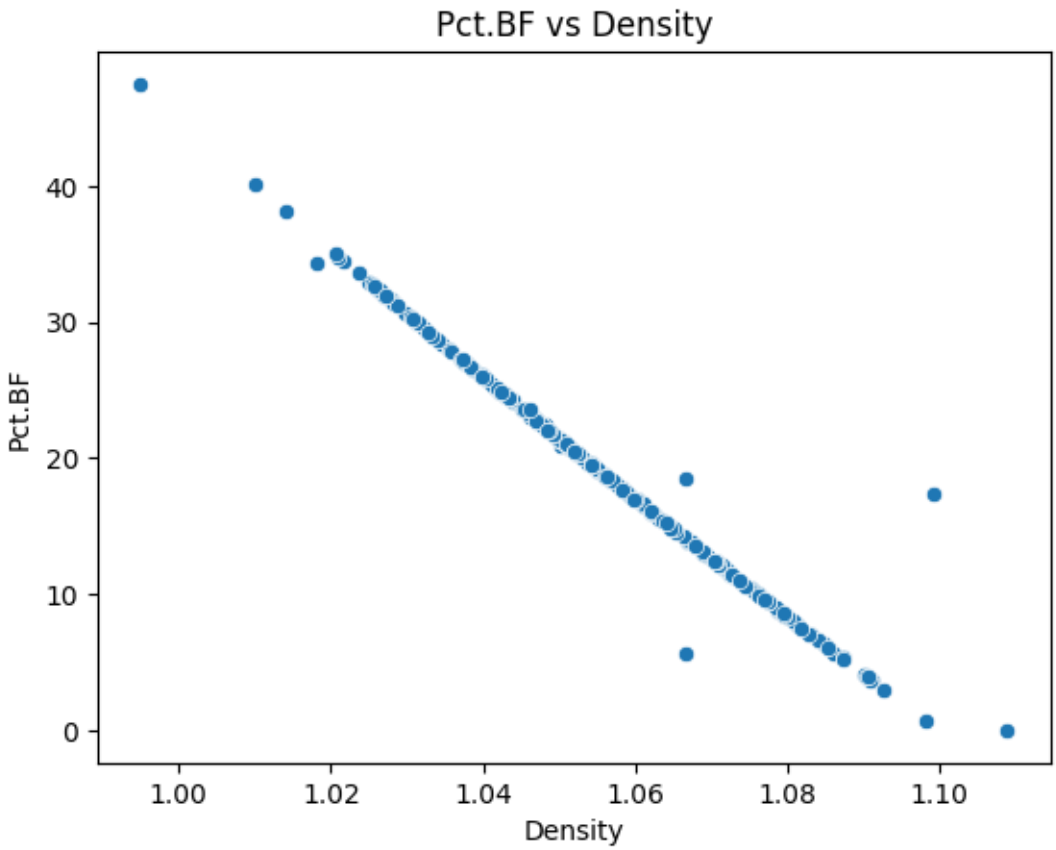


Rys. 1. Wizualizacja macierzy korelacji między zmiennymi

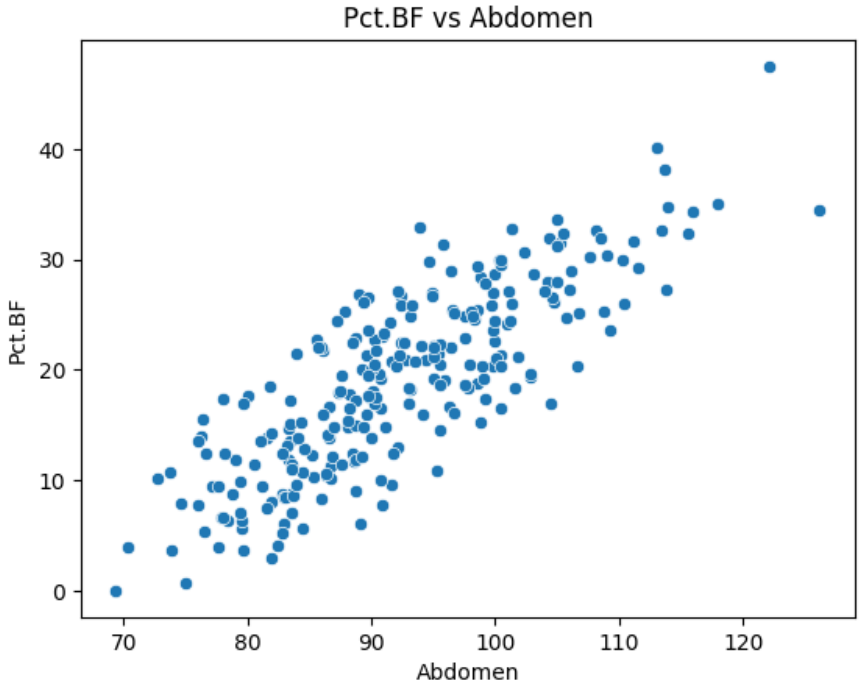
Jak widać na powyższej wizualizacji, istnieje silna korelacja między zmienną zależną – poziomem tkanki tłuszczowej a zmiennymi objaśniającymi – obwodami brzucha i talii. Ponadto, można zauważyć silną korelację między predyktorami – obwodem brzucha i talii, po czym można wnioskować, że dobór zmiennych, np. metodą eliminacji wstecznej, wyeliminuje któryś z nich ze względu na to, że są to bardzo zbliżone miary a zależy nam na unikaniu multikolinearności i redundantnych danych. Warto też zauważyć, że między poziomem tkanki tłuszczowej („Pct. BF") a gęstością („Density") występuje niemalże idealna korelacja ujemna, która oznacza tyle że im większy poziom tkanki tłuszczowej, tym mniejsza gęstość.

I.2.2. Sprawdzenie liniowości na wykresach

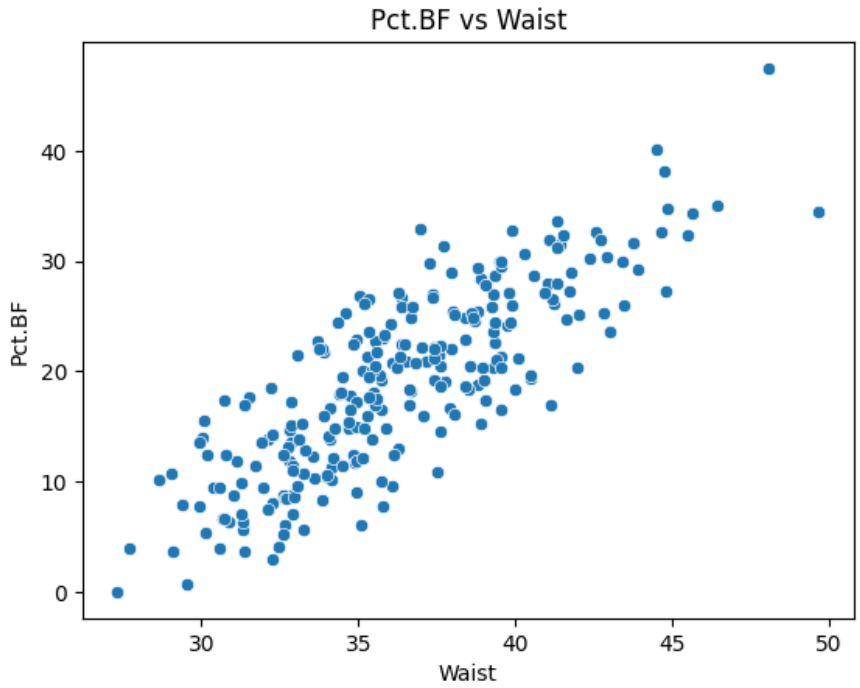
Analiza wykresów rozrzutu danych między zmienną zależną („Pct. BF") a zmiennymi objaśniającymi nie wykazała konkretnego wzorca kształtu związków między nimi. Widać niemalże idealną zależność liniową, silną odwrotną proporcjonalność gęstości w stosunku do poziomu tkanki tłuszczowej. Wykresy pokazujące związek obwodu brzucha czy talii w stosunku do poziomu tkanki tłuszczowej również sugerują zależność liniową. Pokazano także najgorsze wykresy pod kątem sprawdzenia liniowości.



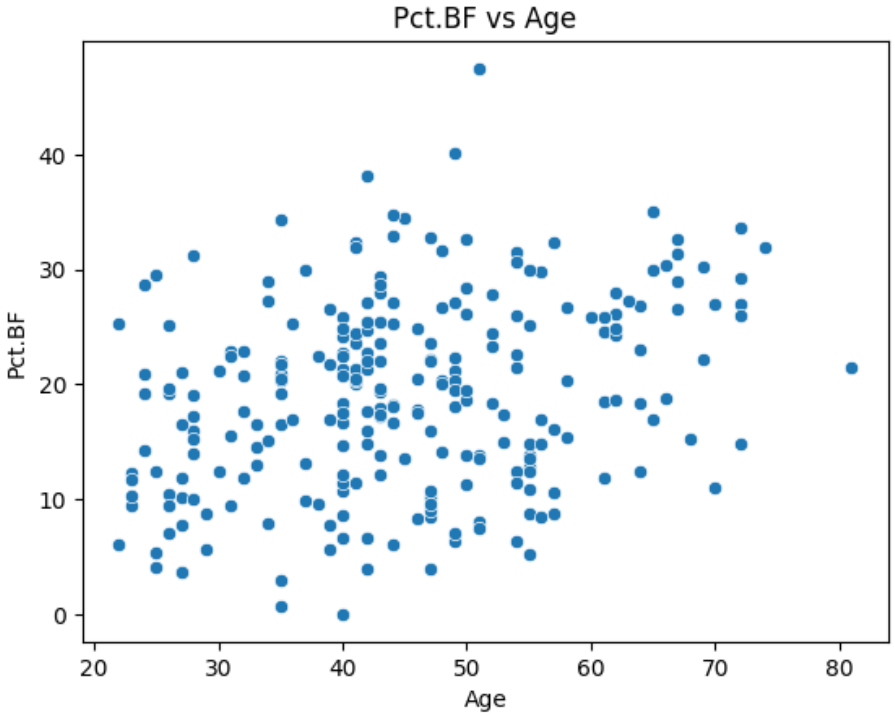
Rys. 2. Wykres gęstości w stosunku do poziomu tkanki tłuszczowej



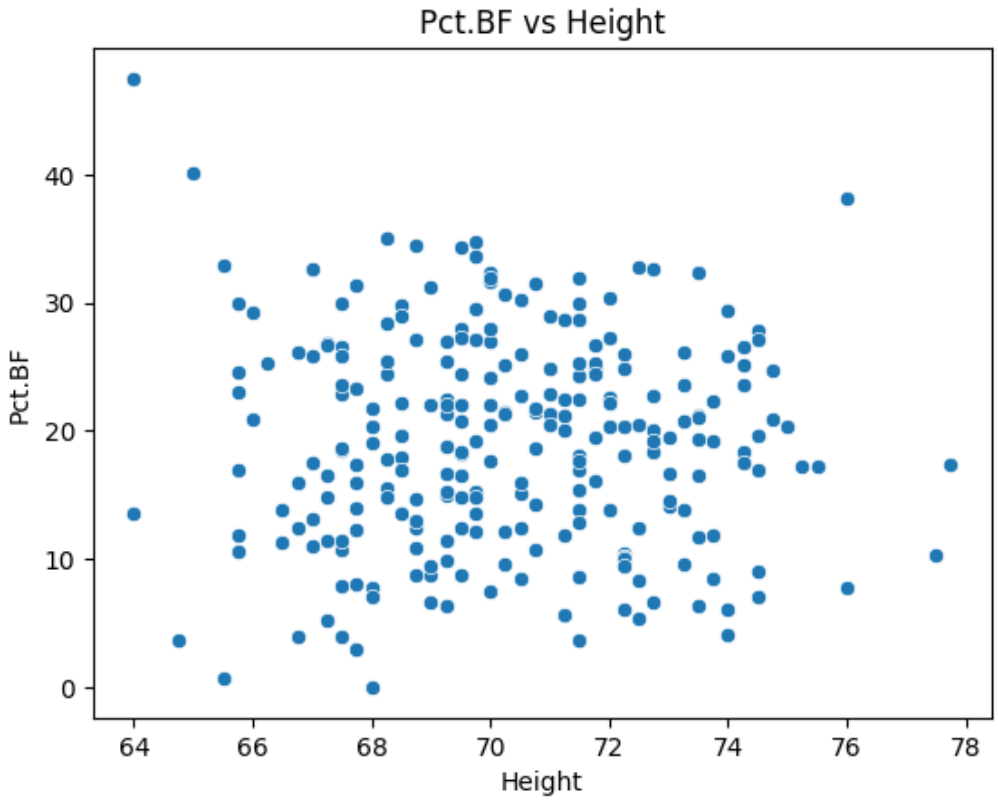
Rys. 3. Wykres punktowy obwodu brzucha w stosunku do poziomu tkanki tłuszczowej



Rys. 4. Wykres punktowy obwodu talii w stosunku do poziomu tkanki tłuszczowej



Rys. 5. Wykres punktowy wieku w stosunku do poziomu tkanki tłuszczowej



Rys. 6. Wykres punktowy wzrostu w stosunku do poziomu tkanki tłuszczowej

I.3. Modele

I.3.1. Model badający poziomu tkanki tłuszczowej na podstawie obwodu brzucha

Z analizy macierzy korelacji wynika występowanie silnej multikolinearności między zmiennymi „Abdomen" i „Waist" (korelacja = 1,0). Aby uniknąć redundancji oraz w celu eksperymentu, zdecydowano o zbudowanie pierwszego modelu opartego wyłącznie o zmienną „Abdomen". Po zbudowaniu modelu, wyznaczono metryki oceny dopasowania modelu do danych.

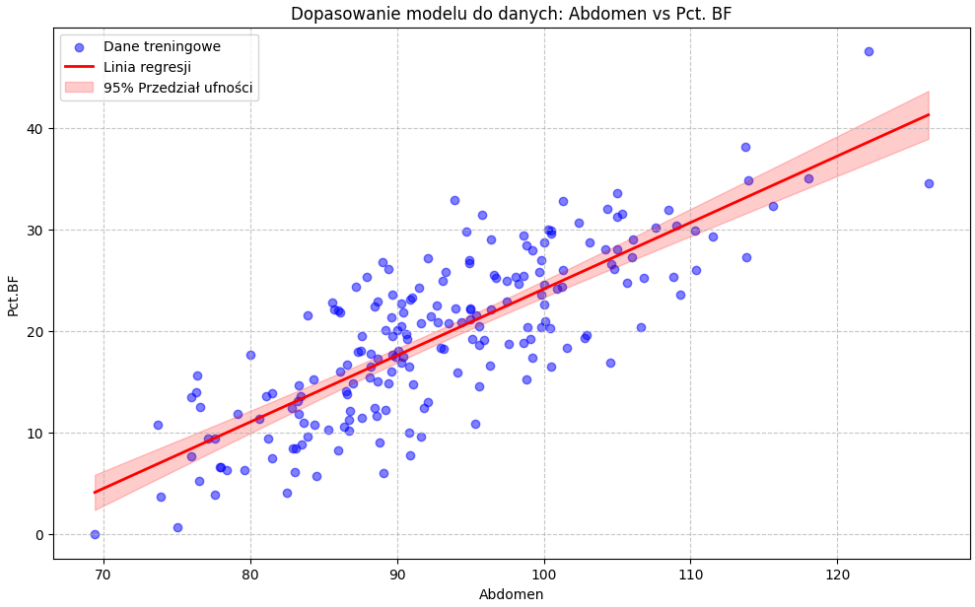
Metryki dopasowania:

$$R^2 = 64{,}5\%$$
$$W_e \approx 26{,}45\%$$

Dokonano także predykcji na danych testowych oraz wyznaczono metryki oceny predykcji powyższego modelu.

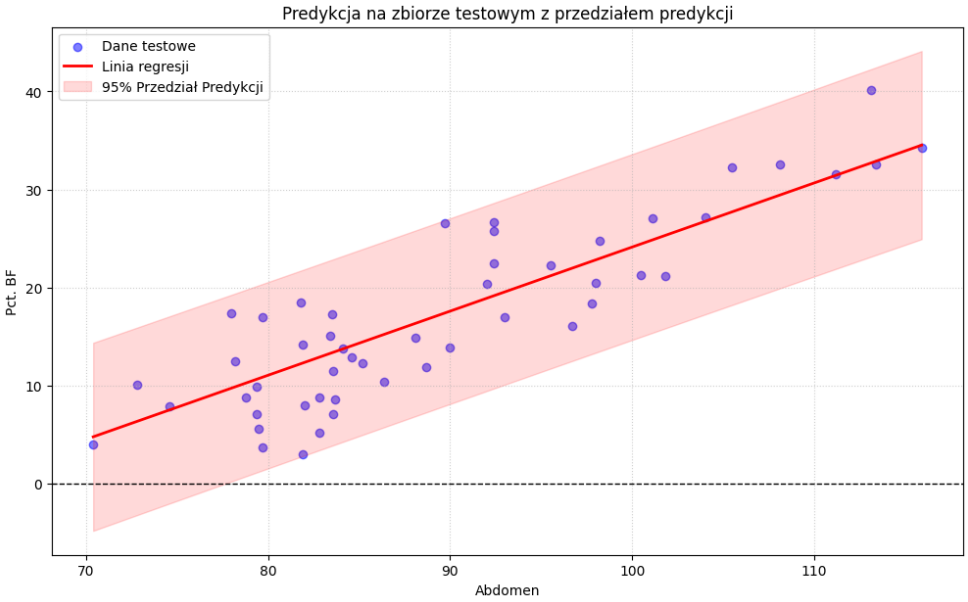
Metryki oceny predykcji:

$$MSE \approx 19{,}53$$
$$MAE \approx 3{,}64$$
$$RMSE \approx 4{,}42$$
$$MAPE \approx 35{,}48\%$$



Rys. 7. Wykres dopasowania modelu regresji liniowej do danych

Model nie dopasowuje się najlepiej do danych treningowych, biorąc pod uwagę współczynnik determinacji oraz współczynnik zmienności losowej, który wynosi ponad 25%. Model cechuje zatem przeciętna zmienność. Wykres dopasowania modelu do danych treningowych wskazuje liniowy trend między obwodem brzucha a poziomem tkanki tłuszczowej. Rozrzut części wartości rzeczywistych poza przedziały ufności może świadczyć o tym, że model zbudowany w oparciu o tą konkretną, jedną, zmienną objaśniającą nie jest najlepszy i najprawdopodobniej istnieją inne cechy, które wpływają na poziom tkanki tłuszczowej oraz które należy uwzględnić, aby poprawić dopasowanie.



Rys. 8. Wykres predykcji modelu w porównaniu z danymi rzeczywistymi (testowymi)

Obliczone wartości błędów MSE, MAE, RMSE, MAPE sugerują, że zdolność modelu do predykcji opartego na tej jednej zmiennej nie jest najlepsza, mimo że wartości testowe zawierają się w przedziałach predykcji dla tego modelu. Średnia kwadratowa różnica między wartościami rzeczywistymi a estymowanymi wynosi 19,53. Przeciętna wielkość błędu bez uwzględniania jego kierunku wynosi 3,64. Wartość procentowa błędów w odniesieniu do rzeczywistych wartości wynosi 35,48%. Analiza metryk i zestawienie danych obserwacji z prognozami sugeruje, że uwzględnienie tylko jednej zmiennej objaśniającej – w tym wypadku „Abdomen", może być niewystarczające do precyzyjnego opisania tak złożonego parametru jakim jest poziom tkanki tłuszczowej, nawet mimo tego, że dane testowe mieszczą się w wyznaczonych przedziałach predykcji. Analiza przedziałów predykcji wykazała istotne ograniczenia tego modelu. Warto zauważyć relatywnie szerokie przedziały predykcji dla poszczególnych rekordów, co może w praktyce nie mieć żadnej wartości biznesowej ze względu na tak duży rozrzut wartości poziomy tkanki tłuszczowej oraz to, że ujemna tkanka tłuszczowa jest biologicznie niemożliwa.

I.3.2. Model badający poziomu tkanki tłuszczowej na podstawie gęstości ciała

Z analizy macierzy korelacji wynikała silna korelacja ujemna między gęstością a poziomem tkanki tłuszczowej (-0,99). Dlatego właśnie zdecydowano się na budowę drugiego modelu, tym razem opartego o zmienną „Density", żeby sprawdzić jak dobrze ten model się sprawdza w naszym konkretnym zadaniu przewidywania poziomu tkanki tłuszczowej. Po zbudowaniu modelu, wyznaczono metryki oceny dopasowania modelu do danych.

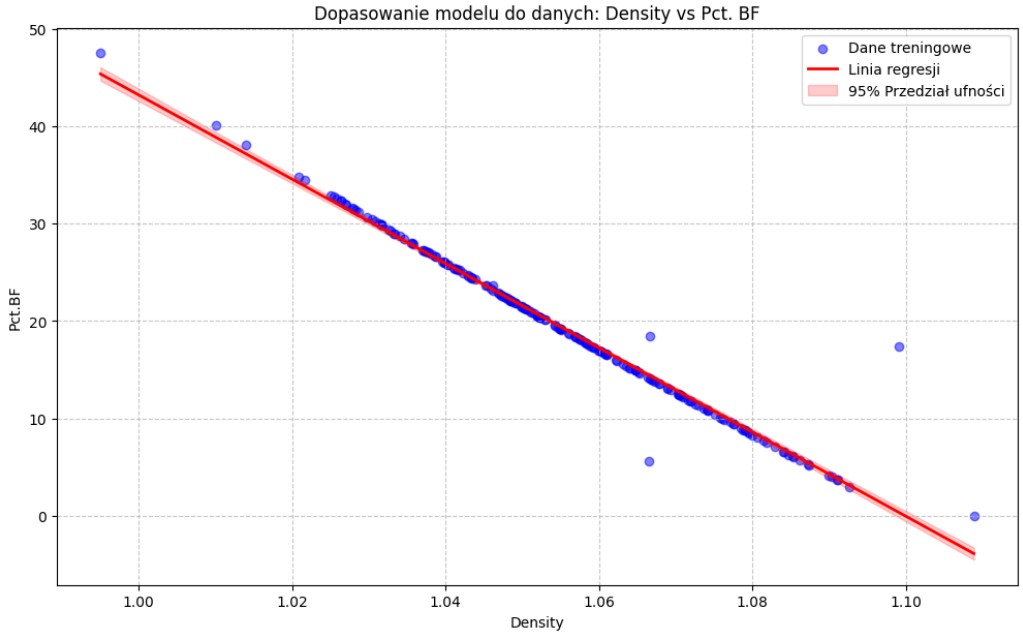
Metryki dopasowania:

$$R^2 = 96{,}9\%$$
$$W_e \approx 8{,}93\%$$

Dokonano także predykcji na danych testowych oraz wyznaczono metryki oceny predykcji powyższego modelu.

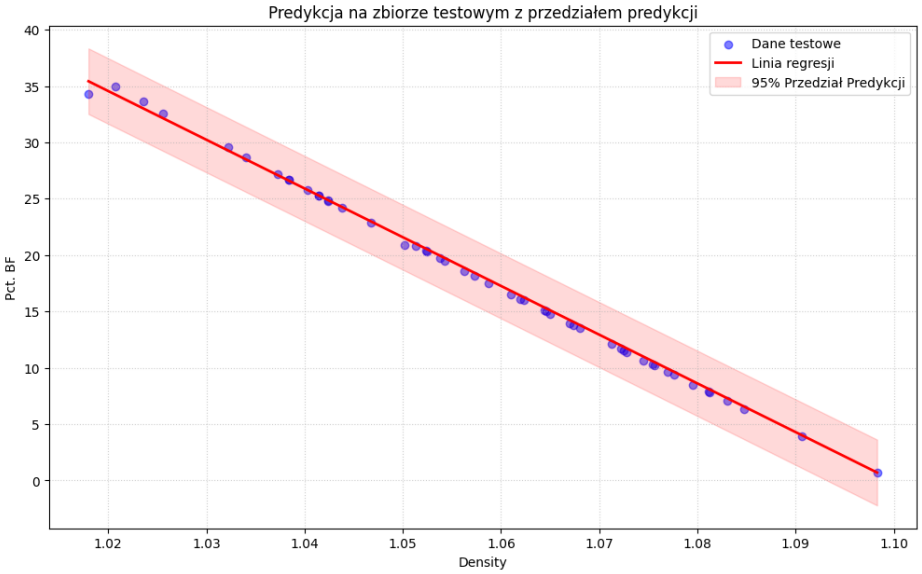
Metryki oceny predykcji:

$$MSE \approx 0{,}11$$
$$MAE \approx 0{,}27$$
$$RMSE \approx 0{,}33$$
$$MAPE \approx 1{,}8\%$$



Rys. 9. Wykres dopasowania modelu regresji liniowej dodanych

Model dobrze, niemalże idealnie, dopasowuje się do danych. Wartości współczynnika determinacji i zmienności losowej na poziomie odpowiednio ok. 97% i 9% świadczą o dobrym dopasowaniu modelu dodanych oraz małej zmienności (współczynnik zmienności losowej jest mniejszy od 25%). Model wyjaśnia ok. 97% zmienności poziomu tkanki tłuszczowej.



Rys. 10. Wykres predykcji modelu w porównaniu z danymi rzeczywistymi (testowymi)

Błędy MSE, MAE, RMSE, MAPE sugerują, że zdolność tego modelu do predykcji jest niemalże idealna. RMSE na poziomie 0,33 oznacza praktycznie pomijalny błąd. MAPE o wartości ok. 1,8% wskazuje, że nasz model myli się o 1,8% prawdziwej wartości. O jakości dopasowania jak i predykcji dowodzą także powyższe wykresy, na którym dane rzeczywiste treningowe i testowe niemalże idealnie pokrywają się z linią regresji.

Zmienna objaśniająca gęstości („Density") zostanie wykluczona z dalszych analiz i modeli, mimo niemalże idealnych metryk. Powodem jest brak realnej wartości biznesowej – gęstość ciała jest niepraktyczna do pomiaru w warunkach, np. gabinetowych. W realnych scenariuszach, takich jak wizyta pacjenta u lekarza, preferujemy łatwo mierzalne cechy, które zachowują wysoką jakość predykcji przy znacznie większej prostocie samego pomiaru.

I.4. Model z wykorzystaniem doboru zmiennych

I.4.1. Model z doborem zmiennych opartym o eliminację wsteczną

Opis metody wraz z uzasadnieniem jej wyboru znajduje się w sekcji I.1.2. Eliminacja wsteczna wyeliminowała zmienne nieistotne statystycznie i zostawiła tylko „Height", „Chest", „Waist", i „Wrist" ze współczynnikami równymi odpowiednio: -0.4964, -0.1999, 2.2885, -1.4077 oraz wyrazem wolnym równym 16.6686. Po zbudowaniu modelu, wyznaczono metryki oceny dopasowania modelu do danych.

Metryki dopasowania:

$$R^2 = 71{,}9\%$$
$$W_e \approx 26{,}98\%$$

Dokonano także predykcji na danych testowych oraz wyznaczono metryki oceny predykcji powyższego modelu.

Metryki oceny predykcji:

$$MSE \approx 19{,}06$$
$$MAE \approx 3{,}50$$
$$RMSE \approx 4{,}37$$
$$MAPE \approx 36{,}3\%$$

Wyznaczone wartości współczynników determinacji oraz zmienności losowej wskazują na lepsze dopasowanie modelu do danych porównując je chociażby z modelem z jedną zmienną objaśniającą – „Abdomen”. Model wyjaśnia ok. 72% zmienności poziomu tkanki tłuszczowej. Współczynnik zmienności losowej na poziomie ok. 27% przekracza próg 25%, wskazując na umiarkowaną precyzję predykcji a zarazem przeciętną zmienność. Wartości błędów MSE, MAE, RMSE są nieco mniejsze od tych dla modelu ze zmienną „Abdomen”. Błąd MAPE jest nieco większy i wskazuje, że model myli się o ok. 36% prawdziwej wartości. Wartości p-value dla każdej ze zmiennych wskazują na to, że są one istotne statystycznie ze względu na wartość mniejszą niż 0,05.

I.5. Model z wykorzystaniem innych metod doboru zmiennych

W ramach realizacji kolejnego punktu polecenia, w celu budowy modeli z innymi zmiennymi objaśniającymi, zdecydowano na przetestowanie innych metod selekcji zmiennych.

I.5.1. Model z doborem zmiennych opartym o regularyzację Lasso

Opis metody znajduje się w sekcji I.1.2. Eliminacja metodą regularyzacji Lasso wyeliminowała (wyzerowała) zmienne nieistotne statystycznie i zostawiła tylko „Age”, „Height”, „Neck”, „Abdomen”, „Waist” i „Wrist” ze współczynnikami równymi odpowiednio: 0.0421, -0.2542, -0.3146, -9.742e+04, 2.475e+05, -1.7323 oraz wyrazem wolnym równym 3.2553. Po zbudowaniu modelu, wyznaczono metryki oceny dopasowania modelu do danych.

Metryki dopasowania:

$$R^2 = 74{,}8\%$$
$$W_e \approx 22{,}83\%$$

Dokonano także predykcji na danych testowych oraz wyznaczono metryki oceny predykcji powyższego modelu.

Metryki oceny predykcji:

$$MSE \approx 19{,}83$$
$$MAE \approx 3{,}63$$
$$RMSE \approx 4{,}45$$
$$MAPE \approx 21{,}4\%$$

Wyznaczone wartości współczynników determinacji oraz zmienności losowej wskazują na lepsze dopasowanie modelu do danych porównując je chociażby z modelem ze zmiennymi wyznaczonymi metodą eliminacji wstecznej (różnica współczynnika determinacji o ok. 2,9 pp.). Model wyjaśnia ok. 75% zmienności poziomu tkanki tłuszczowej. Współczynnik zmienności losowej na poziomie ok. 23% mieści się w przedziale od 0% do 25%, co świadczy o dobrym dopasowaniu modelu do danych oraz małej zmienności. Warto zaznaczyć, że jest to także lepszy wynik w porównaniu do modelu z metodą eliminacji wstecznej. Błędy MSE, RMSE, MAE są nieco większe niż te dla modelu z Model z doborem zmiennych opartym o eliminację wsteczną, natomiast błąd MAPE naszego modelu jest mniejszy, co świadczy o tym, że nasz aktualny model myli się tylko o ok. 21,4% prawdziwej wartości poziomu tkanki tłuszczowej, co jest wynikiem lepszym od naszego poprzedniego modelu. Świadczy to o lepszej jakości predykcji modelu. Metoda Lasso skutecznie zredukowała liczbę zmiennych objaśniających do 6, poprawiając przy tym samym wartość współczynnika zmienności losowej o ok. 4,15 punktu procentowego. Należy jednak wziąć pod uwagę to, że w naszym modelu mamy zarówno zmienną „Abdomen”, jak i „Waist”, które są ze sobą silnie skorelowane, dlatego problem multikolinearności wymaga dalszych działań.

I.5.2. Model z doborem zmiennych opartym o Forward Feature Selection

Opis metody znajduje się w sekcji I.1.2. Eliminacja metodą Forward Feature Selection wyeliminowała zmienne nieistotne statystycznie i zostawiła tylko „Waist”, „Wrist”, „Weight”, „Forearm”, „Neck”, „Age”, „Bicep”, „Ankle” ze współczynnikami równymi odpowiednio: 2.3549, -1.9110, -0.0868, 0.3338, -0.3534, 0.0440, 0.2323, 0.2538 oraz wyrazem wolnym równym -27.8583. Po zbudowaniu modelu, wyznaczono metryki oceny dopasowania modelu do danych.

Metryki dopasowania:

$$R^2 = 75{,}4\%$$
$$W_e \approx 23{,}4\%$$

Dokonano także predykcji na danych testowych oraz wyznaczono metryki oceny predykcji powyższego modelu.

Metryki oceny predykcji:

$$MSE \approx 19{,}87$$
$$MAE \approx 3{,}63$$
$$RMSE \approx 4{,}46$$
$$MAPE \approx 21{,}3\%$$

Wyznaczona wartość współczynnika determinacji wskazuje na lepsze dopasowanie modelu do danych w porównaniu z poprzednimi modelami (różnica o 0,6 pp.). Wartość współczynnika zmienności losowej jest nieco wyższa od wartości tego współczynnika dla poprzedniego rozpatrywanego modelu, ale nadal wskazuje na dobre dopasowanie modelu do danych, gdyż wartość tego współczynnika mieści się w przedziale od 0% do 25%, co świadczy także o małej zmienności. Model wyjaśnia ok. 75% zmienności poziomu tkanki tłuszczowej. Błędy MSE, RMSE, MAE są nieco większe niż te dla modelu z Model z doborem zmiennych opartym o regularyzację Lasso, natomiast błąd MAPE naszego modelu jest mniejszy, co świadczy o tym, że nasz aktualny model myli się tylko o ok. 21,3% prawdziwej wartości poziomu tkanki tłuszczowej, co jest wynikiem lepszym od naszego poprzedniego modelu. Świadczy to o lepszej jakości predykcji modelu. Metoda doboru zmiennych w przód skutecznie zredukowała liczbę zmiennych objaśniających do 8, jednakże wartość współczynnika zmienności losowej zwiększyła się o ok. 0,57 punktu procentowego.

Wnioski

W wyniku analizy różnych modeli regresji liniowej służących do predykcji poziomu tkanki tłuszczowej („Pct. BF") stwierdzono, że żaden z nich nie jest idealny.
Modele jednoczynnikowe („Abdomen vs Pct. BF", „Density vs Pct. BF") wykazują skrajnie niskie i wysokie wyniki dopasowania modelu do danych. Model z „Abdomen" wykazuje się najgorszym dopasowaniem do danych a model z „Density" niemalże idealnym oraz prawie idealnymi metrykami oceny predykcji, ale ma on brak praktycznej użyteczności pomiarowej wyklucza go z realnym zastosowań biznesowych.

Lp.	Metoda doboru zmiennych	R^2	W_e	$MAPE$	Liczba zmiennych
1.	Backward Elimination	71,9%	26,98%	36,3%	4
2.	Regularyzacja Lasso	74,8%	22,83%	21,4%	6
3.	Forward Feature Selection	75,4%	23,40%	21,3%	8

Tab. 1. Porównanie modeli wieloczynnikowych

Metoda Lasso i Forward Feature Selection osiągają najlepsze metryki oceny predykcji, jednakże nie jesteśmy w stanie wybrać najlepszego modelu, gdyż różnice są minimalne a każda z metod wybrała inne zmienne objaśniające. Oba modele charakteryzuje dobre dopasowanie do danych, na podstawie wartości współczynników determinacji oraz zmienności losowej, jednak tutaj także te różnice są relatywnie małe. Należy wziąć pod uwagę to, że nawet te najlepsze modele (o najmniejszej wartości MAPE) mogą zawieść w rzeczywistych warunkach, gdzie kluczowymi czynnikami są te, niemierzalne w danych takie jak chociażby genetyka danej osoby, jej postura, dieta, aktywność fizyczna czy chociażby wiek. Dlatego właśnie złożoność biologiczna człowieka predestynuje te modele do roli raczej wspomagającej niżeli zastępczej.

Bibliografia

[1] https://obliczeniastatystyczne.pl/wspolczynnik-zmiennosci/
[2] https://miroslawmamczur.pl/czym-jest-wybor-zmiennych-feature-selection-16-metod-ktore-warto-znac/
[3] https://exploration.stat.illinois.edu/learn/Feature-Selection/Forward-Selection-Algorithm/
[4] https://www.geeksforgeeks.org/machine-learning/what-is-lasso-regression/
[5] https://medium.com/@lomashbhuva/lasso-regression-l1-regularization-explained-with-practical-examples-a2560a784af2
[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html
[7] https://www.datacamp.com/tutorial/adjusted-r-squared

Spis rysunków

Spis tabel

Załączniki

Lab-1-Zad-2-Michał-Ślęzak-Szymon-Oleśkiewicz.ipynb

Lab-1-Zad-2-Michał-Ślęzak-Szymon-Oleśkiewicz.py

https://colab.research.google.com/drive/1lFh2jkhI_78_rIkMJ8FfvQR8xBIFkZDQ?usp=sharing