**WOJSKOWA  AKADEMIA  TECHNICZNA**

**im. Jarosława Dąbrowskiego**

**WYDZIAŁ  CYBERNETYKI**

**INSTYTUT INŻYNIERII SYSTEMÓW**

SPRAWOZDANIE

z przedmiotu Metody Eksploracji Danych

Zadanie 1.2



**GRUPA WCY23IJ2S1**

|  |  |
| --- | --- |
| **Autorzy:** | **Prowadzący:** |
| **Zbigniew PŁATKOWSKI**  **Martyna FITA** | **dr inż. Romuald Hoffmann** |

**Warszawa 2025/2026**

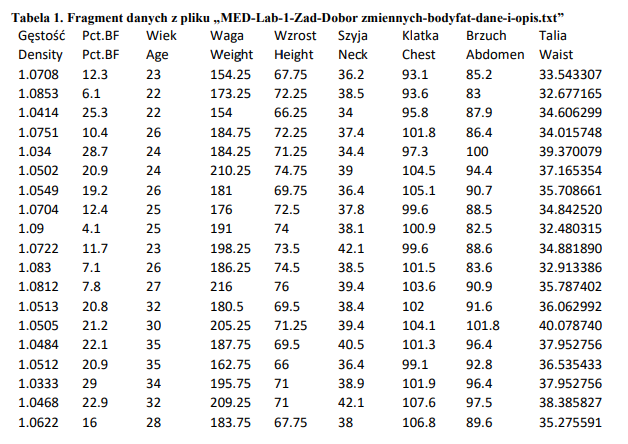
# Treść zadania

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

# Dane

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”



Obraz 1. Fragment danych z pliku

Dane podzielono na zbiór uczący - 200 obserwacji (80%) i zbiór testowy – 50 obserwacji (20%).

# Podstawy teoretyczne

W analizie zależności ilościowych powszechnie stosuje się modele regresji liniowej, które pozwalają na ocenę wpływu poszczególnych zmiennych objaśniających na zmienną objaśnianą. Ze względu na dużą liczbę zmiennych konieczne jest zastosowanie procedur ich selekcji, takich jak eliminacja wsteczna oparta na istotności statystycznej parametrów czy analiza korelacji i współliniowości. Metody te umożliwiają uproszczenie modelu, ograniczenie redundancji informacji oraz zmniejszenie ryzyka przeuczenia.

W zadaniu wykorzystano model regresji liniowej, który zakłada, że zmienna objaśniana zależy liniowo od zmiennych objaśniających , , … , i przyjmuje postać:

, gdzie - wyraz wolny, - współczynniki regresji

Dla danych empirycznych zależność ta zwykle nie jest spełniona dokładnie, dlatego wyznaczamy parametry tak, by model możliwie najlepiej przybliżał obserwacje.

W tym celu stosuje się metodę najmniejszych kwadratów (MNK), która minimalizuje sumę kwadratów reszt:

W szczególności, dla jednej zmiennej () dostajemy postać prostej:

W przypadku modeli nieliniowych estymacja parametrów może być utrudniona. Dlatego często stosuje się linearyzację, polegającą na odpowiednim przekształceniu modelu do postaci liniowej względem parametrów. Pozwala to na wykorzystanie metod regresji liniowej, takich jak metoda najmniejszych kwadratów opisana powyżej.

Powyżej przedstawiono jawne (analityczne) wzory estymatorów metody najmniejszych kwadratów dla modelu regresji liniowej z jedną zmienną objaśniającą, które można wykorzystać do ręcznego policzenia:

**Realizacja prognoz**

Program gretl pozwala na tworzenie modeli regresji liniowej i wykorzystuje je do prognozowania wartości zmiennej objaśnianej. Po zbudowaniu modelu, określającego zależność zmiennej zależnej od zestawu zmiennych niezależnych, gretl może obliczać przewidywane wartości y dla nowych obserwacji. W praktyce oznacza to, że do równania regresji podstawiane są kolejne wartości zmiennych objaśniających, a na tej podstawie program wyznacza odpowiadające im wartości zmiennej objaśnianej.

Istotnym elementem eksploracji danych jest również walidacja modeli na niezależnym zbiorze testowym oraz ocena ich jakości przy użyciu miar błędu prognoz, takich jak średni absolutny błąd procentowy (MAPE).

**Średni absolutny błąd procentowy** (ang. mean absolute percentage error, MAPE)

Przy czym oznacza liczbę punktów (obserwacji), dla których wyznaczono prognozę, czyli liczbę par .

Miara MAPE informuje, o ile procent średnio różnią się wartości prognozowane od wartości rzeczywistych. Dzięki temu pozwala w prosty sposób ocenić dokładność prognoz oraz porównywać jakość predykcji uzyskanych z różnych modeli.

W celu uzyskania przejrzystego i statystycznie poprawnego modelu, podjęto proces selekcji cech mający na celu wyłonienie jedynie tych zmiennych, które wnoszą istotną informację do prognozowania zjawiska. Optymalizację zestawu zmiennych objaśniających przeprowadzono w dwóch etapach: za pomocą algorytmu eliminacji wstecznej oraz analizy współliniowości.

W procedurze eliminacji wstecznej istotność zmiennych objaśniających oceniana jest na podstawie testów statystycznych dotyczących istotności parametrów strukturalnych modelu. Dla każdej zmiennej objaśniającej formułowana jest:

* hipoteza zerowa , mówiąca, że dana zmienna nie ma istotnego wpływu na zmienną objaśnianą,
* hipoteza alternatywna , wskazująca na istotny wpływ tej zmiennej.

Decyzja o odrzuceniu bądź nieodrzuceniu hipotezy zerowej podejmowana jest na podstawie wartości p-value, która oznacza najmniejszy poziom istotności, przy którym hipoteza zerowa zostałaby odrzucona. Jeżeli wartość jest mniejsza od przyjętego poziomu istotności (najczęściej lub ), hipoteza zerowa zostaje odrzucona, a zmienna uznana za statystycznie istotną. W przeciwnym przypadku brak jest podstaw do odrzucenia hipotezy zerowej, co oznacza, że dana zmienna nie jest istotna statystycznie.

W ramach eliminacji wstecznej z modelu usuwa się zmienną objaśniającą o największej wartości p-value, przekraczającej ustalony poziom istotności . Po usunięciu tej zmiennej model estymowany jest ponownie, a procedura testowania istotności powtarzana jest aż do momentu, gdy wszystkie pozostałe zmienne objaśniające okażą się statystycznie istotne**.**

Skąd bierze się wartość p-value?

Wartość *p-value* wynika bezpośrednio z testu t-Studenta służącego do sprawdzania istotności parametru strukturalnego . Dla każdej zmiennej objaśniającej obliczana jest statystyka testowa

Która – przy spełnieniu hipotezy zerowej – ma rozkład t-Studenta o stopniach swobody.

Na podstawie tej statystyki wyznacza się *p-value* jako prawdopodobieństwo uzyskania wartości statystyki co najmniej tak ekstremalnej jak obserwowana, przy założeniu prawdziwości hipotezy zerowej. Innymi słowy, *p-value* odpowiada najmniejszemu poziomowi istotności , dla którego statystyka znalazłaby się w obszarze krytycznym testu t-Studenta.

Jeżeli , to statystyka przekracza wartość krytyczną , co prowadzi do odrzucenia hipotezy zerowej i uznania parametru za istotny statystycznie.

Drugą zastosowaną metodą doboru zmiennych była eliminacja zmiennych na podstawie analizy korelacji i współliniowości. Metoda ta polega na badaniu zależności pomiędzy zmiennymi objaśniającymi oraz pomiędzy zmiennymi objaśniającymi a zmienną objaśnianą. Jeżeli dwie lub więcej zmiennych objaśniających wykazuje silną wzajemną korelację, oznacza to, że przekazują one podobną informację, co może prowadzić do problemu współliniowości w modelu.

W takim przypadku z modelu eliminuje się jedną z silnie skorelowanych zmiennych, wybierając tę, która wykazuje słabszy związek ze zmienną objaśnianą. Siłę zależności liniowej pomiędzy zmiennymi mierzy się za pomocą współczynnika korelacji liniowej Pearsona, który przyjmuje wartości z przedział [-1, 1]. Wartości bliskie 1 lub -1 oznaczają silną zależność liniową, natomiast wartości bliskie wskazują na brak zależności liniowej. W praktyce analizuje się zarówno korelacje pomiędzy zmiennymi objaśniającymi, jak i korelacje zmiennych objaśniających ze zmienną objaśnianą, co umożliwia identyfikację zmiennych współliniowych i wybór tych, które w większym stopniu wyjaśniają zmienność zmiennej objaśnianej.

Wzór na współczynnik korelacji Pearsona

gdzie oznaczają obserwacje zmiennych i , natomiast i ich średnie arytmetyczne.

Badane zależności

W analizie zostaną zbadane dwie kluczowe zależności. Pierwszą z nich jest, jaki czynnik w największym stopniu wpływa na procent tkanki tłuszczowej. Odpowiedź na to pytanie uzyskano eliminując poszczególne zmienne aż została jedna, która w największym stopniu determinuje badaną cechę. Drugą zależnością, jaką zdecydowaliśmy się zbadać jest, co wpływa na cechę uzyskaną w pierwszym badaniu. Uzasadnienie opisano niżej w sprawozdaniu.

# Co w największym stopniu wpływa na procent tkanki tłuszczowej?

Zastosowano model liniowy z piętnastoma zmiennymi x1-x15 oraz jedną zmienną zależną y – PctBF i wykorzystując metodę eliminacji wstecznej. Wyniki przedstawiono poniżej. Wzór, który otrzymaliśmy:

|  |  |
| --- | --- |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 2. Model ze wszystkimi zmiennymi. Największa wartość p jest dla zmiennej Hip, więc jako pierwsza zostaje usunięta z modelu. | A screenshot of a computer  AI-generated content may be incorrect.  Obraz 3. Model po usunięciu zmiennej Hip. Największa wartość p jest dla zmiennej Height, więc jako następna zostaje usunięta z modelu. |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 4. Model po usunięciu zmiennej Height. Największa wartość p jest dla zmiennej Thigh, więc jako następna zostaje usunięta z modelu. | A screenshot of a computer  AI-generated content may be incorrect.  Obraz 5. Model po usunięciu zmiennej Thigh. Największa wartość p jest dla zmiennej Forearm, więc jako następna zostaje usunięta z modelu. |

|  |  |
| --- | --- |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 6. Model po usunięciu zmiennej Forearm. Największa wartość p jest dla zmiennej Wrist, więc jako następna zostaje usunięta z modelu. | A screenshot of a computer  AI-generated content may be incorrect.  Obraz 7. Model po usunięciu zmiennej Wrist. Największa wartość p jest dla zmiennej Knee, więc jako następna zostaje usunięta z modelu. |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 8. Model po usunięciu zmiennej Thigh. Największa wartość p jest dla zmiennej Forearm, więc jako następna zostaje usunięta z modelu. | Obraz 9. Model po usunięciu zmiennej Waist. Największa wartość p jest dla zmiennej Neck, więc jako następna zostaje usunięta z modelu. |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 10. Model po usunięciu zmiennej Neck. Największa wartość p jest dla zmiennej Abdomen, więc jako następna zostaje usunięta z modelu. | Obraz 11. Model po usunięciu zmiennej Abdomen. Największa wartość p jest dla zmiennej Ankle, więc jako następna zostaje usunięta z modelu. |
| Obraz 12. Model po usunięciu zmiennej Ankle. Największa wartość p jest dla zmiennej Weight, więc jako następna zostaje usunięta z modelu. | Obraz 13. Model po usunięciu zmiennej Weight. Największa wartość p jest dla zmiennej Bicep, więc jako następna zostaje usunięta z modelu. |
| Obraz 14. Model po usunięciu zmiennej Bicep. Dla żadnej ze zmiennych wartość p nie przekracza przyjętego poziomu istotności, więc na tym etapie możemy zakończyć eliminację wsteczną. | |

A screenshot of a graph

AI-generated content may be incorrect.

Obraz 15. Prognoza oparta na zbiorze testowym (201 - 250), z wykorzystaniem modelu ze zmiennymi objaśniającymi Chest, Age, Density

A screenshot of a computer

AI-generated content may be incorrect.

Obraz 16. Miary dokładności prognoz modelu ze zmiennymi objaśniającymi Chest, Age, Density

Dla otrzymanego modelu zastosowano metodę eliminacji zmiennych na podstawie korelacji i kolinearności i otrzymano następujące wyniki:

|  |  |
| --- | --- |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 17. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Chest a Density. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą PctBF ma zmienna Chest, więc zostaje usunięta. | Obraz 18. Model po usunięciu zmiennej Chest. |
| A screenshot of a computer  AI-generated content may be incorrect.  Obraz 19. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Age a Density. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą PctBF ma zmienna Age, więc zostaje usunięta. | Obraz 20. Model po usunięciu zmiennej Age. |

Po wykonaniu tej metody, w modelu pozostaje zmienna objaśniająca, która ma największy wpływ na zmienną objaśnianą PctBF, czyli zmienna Density. Zależność ta wyraża się wzorem:

Korelację pomiędzy tymi zmiennymi obliczono w programie gretl i uzyskano następujący wynik:

A screenshot of a computer

AI-generated content may be incorrect.

Obraz 21. Korelacja zmiennych Density i PctBF

A screen shot of a computer

AI-generated content may be incorrect.

Obraz 22. Wykres przedstawiający zależność między Density a PctBF

A screen shot of a graph

AI-generated content may be incorrect.

Obraz 23. Prognoza oparta na zbiorze testowym (201 - 250), z wykorzystaniem modelu ze zmienną objaśniającą Density

A screenshot of a computer

AI-generated content may be incorrect.

Obraz 24. Miary dokładności prognoz modelu ze zmienną objaśniającą Density

W prognozie dla modelu z jedną zmienną objaśniającą MAPE wynosi 2,1285, a dla modelu z trzema zmiennymi 4,3804. Pokazuje to, że model uzyskany po zastosowaniu metody eliminacji zmiennych na podstawie korelacji i kolinearności lepiej poradził sobie z prognozą na zbiorze testowym.

# Co w największym stopniu wpływa na gęstość?

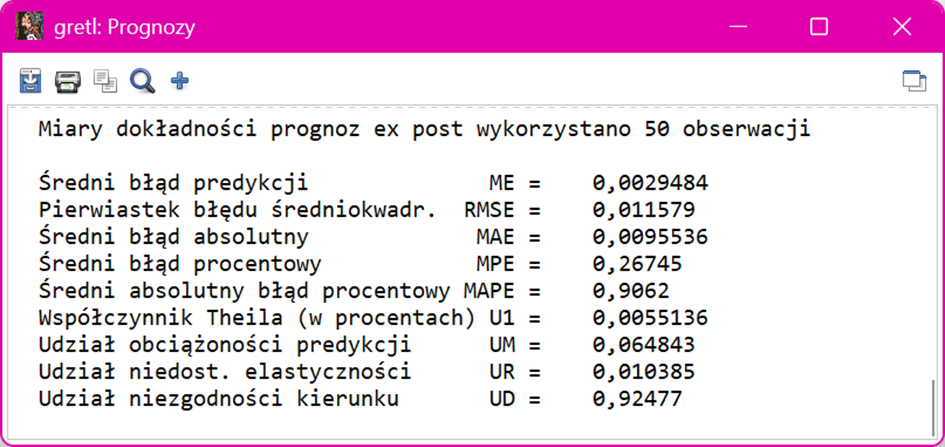
Zastosowano model liniowy z czternastoma zmiennymi x1-x14 oraz jedną zmienną zależną y – Density i wykorzystano metodę eliminacji wstecznej, czego wyniki przedstawiono poniżej. Początkowy wzór na Density przed eliminacją zmiennych:

|  |  |
| --- | --- |
| Obraz 25. Model ze wszystkimi zmiennymi. Największa wartość p jest dla zmiennej Waist, więc jako pierwsza zostaje usunięta z modelu. | Obraz 26. Model po usunięciu zmiennej Waist. Największa wartość p jest dla zmiennej Chest, więc jako następna zostaje usunięta z modelu. |
| Obraz 27. Model po usunięciu zmiennej Waist. Największa wartość p jest dla zmiennej Knee, więc jako następna zostaje usunięta z modelu. | Obraz 28. Model po usunięciu zmiennej Knee. Największa wartość p jest dla zmiennej Height, więc jako następna zostaje usunięta z modelu. |

|  |  |
| --- | --- |
| Obraz 29. Model po usunięciu zmiennej Height. Największa wartość p jest dla zmiennej Bicep, więc jako następna zostaje usunięta z modelu. | Obraz 30. Model po usunięciu zmiennej Bicep. Dla żadnej ze zmiennych wartość p nie przekracza przyjętego poziomu istotności, więc na tym etapie możemy zakończyć eliminację wsteczną. |



Obraz 31. Prognoza oparta na zbiorze testowym (201 - 250), z wykorzystaniem modelu ze zmiennymi objaśniającymi Age, Weight, Neck, Abdomen, Hip, Thigh, Ankle, Forearm, Wrist



Obraz 32. Miary dokładności prognoz modelu ze zmiennymi objaśniającymi Age, Weight, Neck, Abdomen, Hip, Thigh, Ankle, Forearm, Wrist

Dla otrzymanego modelu zastosowano metodę eliminacji zmiennych na podstawie korelacji i kolinearności i otrzymano następujące wyniki:

|  |  |
| --- | --- |
| Obraz 33. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Hip a Weight. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Weight, więc zostaje usunięta. | Obraz 34. Model po usunięciu zmiennej Weight. Program gretl informuje, że wartość p przekracza przyjęty poziom istotności dla zmiennej Ankle, dlatego jako kolejna zostaje usunięta z modelu, zanim przejdziemy do dalszej analizy na podstawie korelacji. |
| Obraz 35. Model po usunięciu zmiennej Ankle. Żadna z wartości p nie przekracza przyjętego poziomu istotności, więc wracamy do analizy korelacji. | |

|  |  |
| --- | --- |
| Obraz 36. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Hip a Thigh. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Thigh, więc zostaje usunięta. | Obraz 37. Model po usunięciu zmiennej Thigh. |
| Obraz 38. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Hip a Abdomen. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Hip, więc zostaje usunięta. | Obraz 39. Model po usunięciu zmiennej Hip. |
| Obraz 40. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Wrist a Neck. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Wrist, więc zostaje usunięta. | Obraz 41. Model po usunięciu zmiennej Wrist. Program gretl informuje, że wartość p przekracza przyjęty poziom istotności dla zmiennej Forearm, dlatego jako kolejna zostaje usunięta z modelu, zanim przejdziemy do dalszej analizy na podstawie korelacji. |

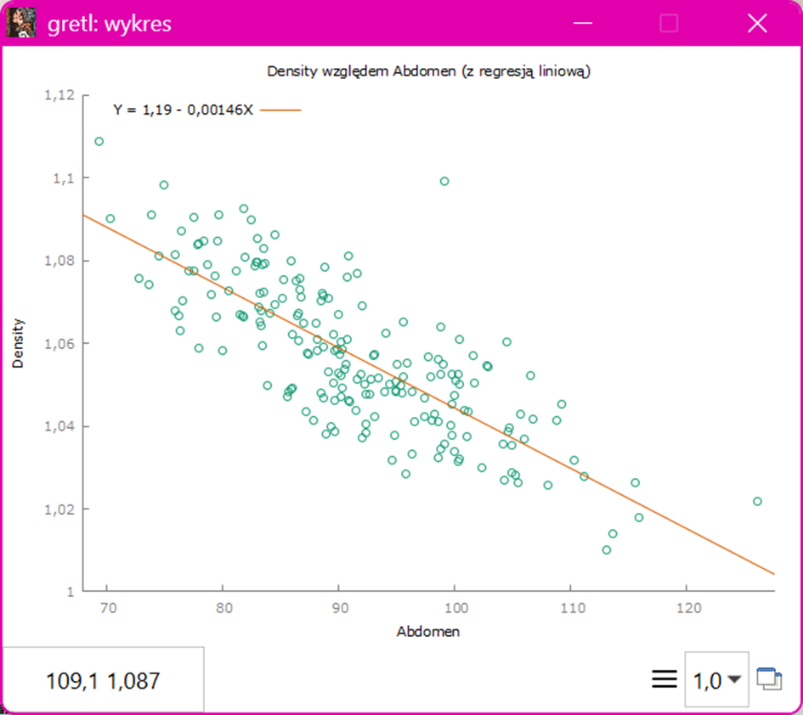
|  |  |
| --- | --- |
| Obraz 42. Model po usunięciu zmiennej Forearm. Żadna z wartości p nie przekracza przyjętego poziomu istotności, więc wracamy do analizy korelacji. | |
| Obraz 43. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Abdomen a Neck. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Neck, więc zostaje usunięta. | Obraz 44. Model po usunięciu zmiennej Neck. |
| Obraz 45. Największa korelacja między zmiennymi objaśniającymi co do wartości bezwzględnej występuje między Abdomen a Age. Spośród tych dwóch mniejszy wpływ na zmienną objaśnianą Density ma zmienna Age, więc zostaje usunięta. | Obraz 46. Model po usunięciu zmiennej Age. |

Po wykonaniu tej metody, w modelu pozostaje zmienna objaśniająca, która ma największy wpływ na zmienną objaśnianą Density, czyli zmienna Abdomen. Wzór po wyeliminowaniu wszystkich zmiennych ma postać:

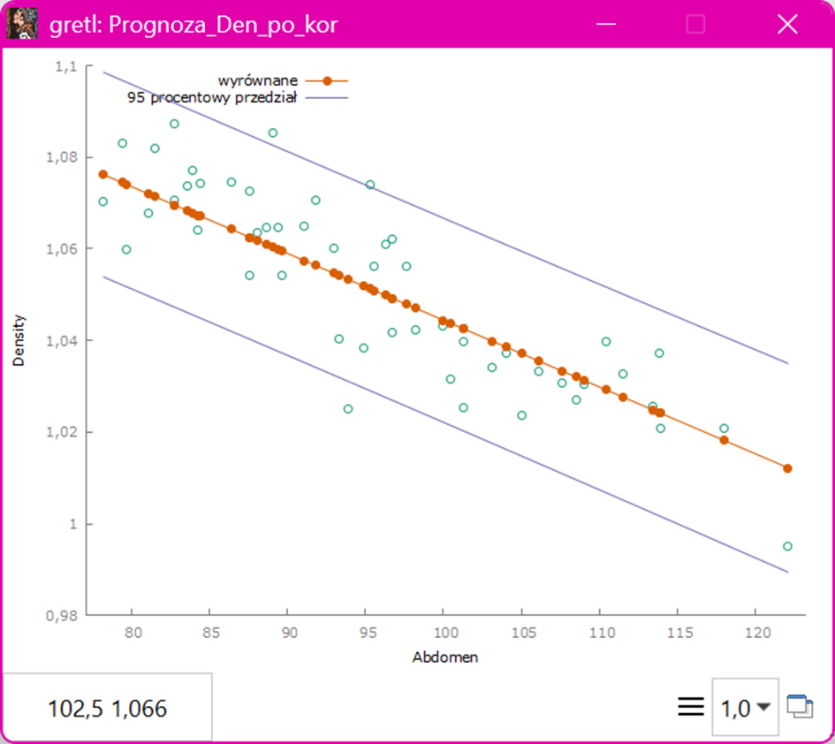
W programie gretl obliczono także korelację pomiędzy tymi zmiennymi i uzyskano:



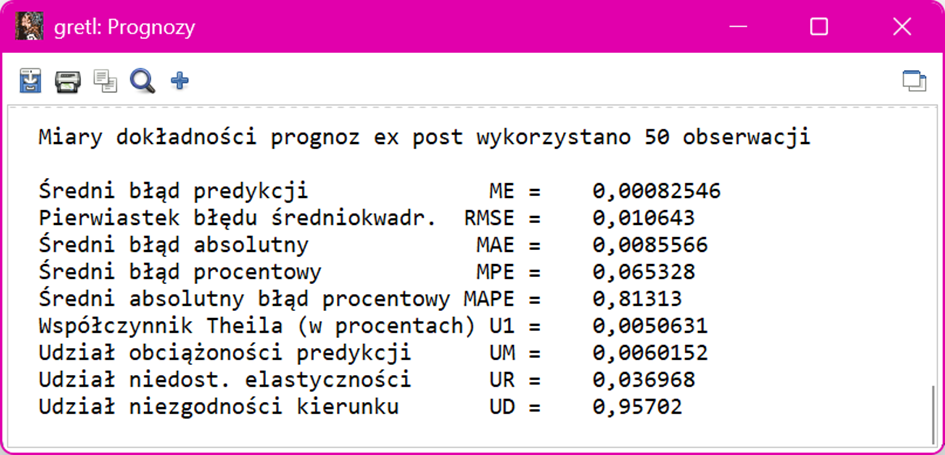
Obraz 47. Korelacja zmiennych Abdomen i Density



Obraz 48. Wykres przedstawiający zależność między Abdomen a Density



Obraz 49. Prognoza oparta na zbiorze testowym (201 - 250), z wykorzystaniem modelu ze zmienną objaśniającą Abdomen



Obraz 50. Miary dokładności prognoz modelu ze zmienną objaśniającą Density

W prognozie dla modelu z jedną zmienną objaśniającą MAPE wynosi 0,81313, a dla modelu z dziewięcioma zmiennymi 0,9062. Pokazuje to, że model uzyskany po zastosowaniu metody eliminacji zmiennych na podstawie korelacji i kolinearności lepiej poradził sobie z prognozą na zbiorze testowym.

# Wnioski

Przeprowadzona analiza wykazała, że najsilniejszy i najbardziej istotny wpływ na procent tkanki tłuszczowej (%bodyfat) ma zmienna Density. Uzyskana silna zależność pomiędzy gęstością ciała a zawartością tłuszczu jest w pełni zgodna z wiedzą fizjologiczną, ponieważ wzrost udziału tkanki tłuszczowej prowadzi do obniżenia gęstości całkowitej ciała. Widać to jednoznacznie we wzorze:

, o nazwie Siri Equation.

Potwierdza to zasadność stosowania Density jako kluczowej zmiennej w modelach predykcyjnych %bodyfat.

Dalsza analiza wykazała, że zmienną najsilniej determinującą gęstość ciała u badanych mężczyzn jest obwód brzucha (Abdomen). Wynik ten ma uzasadnienie biologiczne, ponieważ u mężczyzn tkanka tłuszczowa w dużym stopniu odkłada się w okolicach jamy brzusznej, co istotnie wpływa na proporcje pomiędzy masą tłuszczową a beztłuszczową, a tym samym na gęstość ciała.

Należy jednocześnie zauważyć, że uzyskane rezultaty dotyczą wyłącznie populacji mężczyzn. W przypadku kobiet zależności te mogłyby mieć inny charakter, ze względu na odmienny sposób rozmieszczenia tkanki tłuszczowej. W takiej populacji inne zmienne obwodowe mogłyby mieć większe znaczenie w modelowaniu gęstości ciała oraz procentu tkanki tłuszczowej.