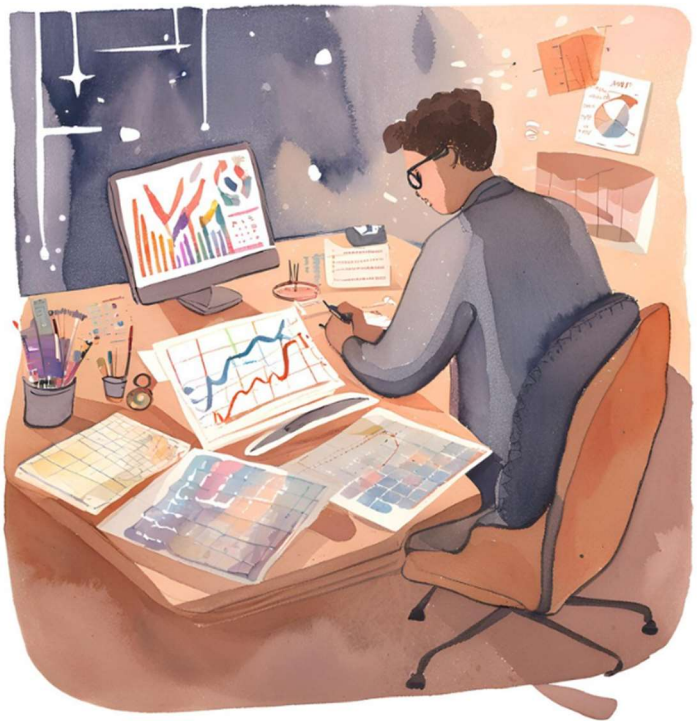
Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Ćwiczenie nr 11
Testy korelacji i modele regresji



POLITECHNIKA BYDGOSKA
im. Jana i Jędrzeja Śniadeckich

POLITECHNIKA BYDGOSKA
Wydział Technologii i Inżynierii Chemicznej

POLITECHNIKA BYDGOSKA
Wydział Medyczny

PRACOWNIA KOMPUTEROWA

Wstęp

Testy korelacji i modele regresji są bardzo często wykorzystywanymi narzędziami w analizie zależności między zmiennymi. Korelacja pozwala ocenić siłę i kierunek powiązań między dwiema zmiennymi, co ma istotne znaczenie w badaniach obserwacyjnych, takich jak związek między parametrami klinicznymi a wynikami leczenia. Modele regresji, z kolei, umożliwiają bardziej zaawansowaną analizę, w tym przewidywanie wyników na podstawie zestawu zmiennych. W praktyce medycznej i badaniach naukowych umiejętność korzystania z tych narzędzi jest bardzo ważna ponieważ pozwala na formułowanie hipotez oraz podejmowanie decyzji diagnostycznych i terapeutycznych opartych na danych.

1. Testy korelacji

a) Współczynnik korelacji Pearsona – mierzy zależność liniową między dwiema zmiennymi liczbowymi; zakłada normalność rozkładu.
   - w języku R: cor() lub cor.test(x, y, method="pearson")
   - w Excelu =PEARSON(tab1,tab2)
   - w Excelu =corr(tablica) liczy macierz korelacji - dla więcej niż 2 zmiennych

b) Korelacja Spearmana – test nieparametryczny, stosowany do badania monotonicznych zależności między zmiennymi, szczególnie gdy dane nie są normalnie rozłożone.
   - w języku R: cor.test(x, y, method = "spearman")

Ograniczenia interpretacji:

- Korelacja Pearsona mierzy tylko liniowy związek między zmiennymi. Jeśli związek nie jest liniowy, wynik może być zaniżony lub wprowadzać w błąd.
- Korelacja Spearmana jest bardziej odporna na odstające obserwacje i lepiej mierzy zależności nieliniowe niż korelacja Pearsona, ale nie mówi nic o dokładnej formie tej zależności (np. liniowej czy nieliniowej).

Wzór na statystykę t-Studenta dla korelacji Pearsona (do policzenia p-value):

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}},$$

gdzie:

- r to współczynnik korelacji Pearsona.
- n to liczba obserwacji (rozmiar próby).

Po obliczeniu wartości t, użyj funkcji liczącej dwustronny rozkład t_Studenta (w Excelu ROZKŁ.T.DS) dla obliczenia dwustronnej p-wartości. Wynik tej funkcji to p-wartość dla hipotezy zerowej, że korelacja w populacji wynosi 0 (brak korelacji). P-value < 0.05 mówi, że odrzucamy H0, więc że korelacja jest istotna statystycznie.

2. Modele regresji

Analiza regresji to metoda statystyczna, która pozwala zrozumieć zależności między zmiennymi. Najczęściej analizowana jest regresja liniowa, w której staramy się opisać zależność liniową między zmienną niezależną (np. wiekiem pacjentów) a zmienną zależną (np. ciśnieniem krwi).

- Regresja liniowa (w języku R: lm()) – stosowana do modelowania zależności liniowych między zmienną zależną a jedną lub większą liczbą zmiennych niezależnych.
- Regresja logistyczna (w języku R: glm(family = binomial)) – używana do przewidywania zmiennej binarnej na podstawie jednej lub wielu zmiennych niezależnych.
- Regresja wielokrotna – rozbudowa regresji liniowej dla wielu zmiennych objaśniających; w języku R analizowana za pomocą lm().

Aby ocenić jakość dopasowania oblicza się współczynnik determinacji R^2. Jest to miara, która mówi, jaka część wariancji zmiennej zależnej jest wyjaśniona przez model regresji. Innymi słowy, mówi jak dobrze zmienne niezależne (np. wiek, styl życia) wyjaśniają zmienność zmiennej zależnej (np. ciśnienia krwi).
Zakres wartości R^2 mieści się w przedziale od 0 do 1.

- $R^2 = 0$: Model nie wyjaśnia żadnej wariancji w danych.
- $R^2 = 1$: Model doskonale wyjaśnia całą wariancję w danych.

Wartość R^2 bliska 1 oznacza, że model dobrze dopasowuje się do danych i wyjaśnia dużą część zmienności zmiennej zależnej.

3. Korelacja vs kowariancja.

Kowariancja i korelacja to dwie miary statystyczne opisujące zależność między dwiema zmiennymi. Chociaż są ze sobą powiązane, różnią się w interpretacji, skali i zastosowaniu. Kowariancja mierzy kierunek i siłę związku liniowego między dwiema zmiennymi. Mówi, czy wartości jednej zmiennej, X, rosną (lub maleją) w sposób systematyczny wraz ze wzrostem (lub spadkiem) wartości drugiej zmiennej, Y. Gdy kowariancja jest bliska 0 to jest brak systematycznego związku między zmiennymi.

$$\text{Kowariancja (X, Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Kowariancja zależy od jednostek miary zmiennych, co utrudnia porównywanie jej między różnymi zestawami danych. Np. kowariancja między wzrostem (w cm) a wagą (w kg) ma inne wartości niż między wzrostem (w m) a wagą (w kg).

Korelacja mierzy siłę i kierunek związku liniowego między dwiema zmiennymi, podobnie jak kowariancja, ale dodatkowo jest znormalizowana. Dzięki temu przyjmuje wartości w ustalonym zakresie od -1 do 1, co czyni ją bardziej uniwersalną. Im bliżej wartości 1 lub -1, tym silniejszy jest związek.

$$r = \frac{\text{Kowariancja (X, Y)}}{\sigma_X \cdot \sigma_Y},$$

gdzie σ_X i σ_Y to odchylenia standardowe zmiennych X i Y.

Kiedy używamy kowariancji?

- Zrozumienie kierunku zależności: Kowariancja wskazuje, czy zmienne są dodatnio (gdy wartości jednej zmiennej rosną, druga też rośnie) czy ujemnie (gdy jedna rośnie, druga maleje) skorelowane.
- W analizach wstępnych: Używana do podstawowej oceny współzależności zmiennych przed normalizacją danych.

- W algorytmach statystycznych: Kowariancja jest kluczowa w macierzy kowariancji, która opisuje współzmienność wielu zmiennych jednocześnie (np. w analizie PCA - analizy głównych składowych).
- Brak potrzeby standaryzacji: Używana, gdy różnice w jednostkach miary między zmiennymi są istotne, np. jeśli chcemy zobaczyć rzeczywiste wartości zmienności.

Przykład zastosowania: Chcesz zobaczyć, jak liczba godzin nauki zmienia się wraz z liczbą zdobytych punktów na egzaminie, a skala jednostek nie przeszkadza.

Kiedy używamy korelacji?

- Porównanie siły zależności: Korelacja jest znormalizowaną wersją kowariancji, dzięki czemu przyjmuje wartości w zakresie od -1 do 1. Używana do oceny, jak silnie i w jakim kierunku dwie zmienne są skorelowane.
- Porównanie zmiennych w różnych jednostkach: Gdy zmienne mają różne jednostki miary lub różne zakresy, korelacja pozwala na porównanie ich zależności w sposób bezwymiarowy.
- W praktyce: Częściej używana, ponieważ jej interpretacja jest prostsza i bardziej intuicyjna (np. "silna dodatnia korelacja").
- Do budowy modeli: Używana w analizie regresji i do oceny, które zmienne są istotnie powiązane.

Przykład zastosowania: jeśli chcemy sprawdzić, czy istnieje zależność między temperaturą a sprzedażą lodów, ale te zmienne mają różne jednostki (stopnie Celsjusza i liczba sprzedanych sztuk).

Cel

Celem ćwiczenia jest zrozumienie podstawowych pojęć związanych z korelacją i regresją, takich jak współczynnik korelacji, wartość p, czy współczynnik determinacji R^2. Studenci nabędą praktyczne umiejętności w przeprowadzaniu testów korelacji i budowaniu modeli regresji liniowej za pomocą wybranego oprogramowania statystycznego (np. Excel, R). Nauczą się jak interpretować wyniki analiz w kontekście medycznym. Celem jest również rozwinięcie krytycznego myślenia poprzez ocenę poprawności zastosowania metod statystycznych i ich ograniczeń w analizach medycznych.

Przebieg ćwiczenia

Pobierz plik dane_cw11.xlsx (link poda prowadzący), zawierający dane: wiek, BMI, cholesterol oraz ciśnienie, na których wykonasz następujące ćwiczenia:

1. Obliczenie współczynnika korelacji w Excelu

Otwórz plik dane_cw11.xlsx.

- Oblicz współczynnik korelacji między BMI a cholesterolem. W tym celu zaznacz pustą komórkę i wpisz formułę:
  =PEARSON(zakres1; zakres2)
  gdzie zakres1 to zakres danych pierwszej zmiennej (np. B2:B21), a zakres2 to zakres danych drugiej zmiennej (np. C2:C21).
- Oblicz p-value dla współczynnika korelacji. W tym celu najpierw policz wartość t w następujący sposób:

=G5*PIERWIASTEK((ILE.NIEPUSTYCH(B2:B21)-2)/(1-G5^2))
Uwaga: zakładam, że w komórce G5 policzono współczynnik korelacji - dostosuj ten adres komórki do swojego pliku.
Następnie policz p-value ze wzoru:
=ROZKŁ.T.DS(MODUŁ.LICZBY(G6);ILE.NIEPUSTYCH(B2:B21)-2)
Uwaga: zakładam, że w komórce G6 policzono wartość t z poprzedniego polecenia
Zinterpretuj wynik.

- Oblicz współczynnik korelacji dla reszty zmiennych – stwórz w tym celu macierz korelacji:
  - w Office o365 za pomocą formuły: =CORR(A2:E21)
  - w innych wersjach Excela za pomocą pakietu Analysis ToolPak (Dane->Analiza danych->Korelacja). Zaznacz wszystkie kolumny (wraz z nagłówkami), między którymi chcesz policzyć koleracje. Zaznacz też „Tytuły w pierwszym wierszu".

Między którymi zmiennymi jest najmniejsza a między którymi największa korelacja?

2. Wykres punktowy z linią trendu w Excelu

a) Linia trendu
Zaznacz dane dwóch wybranych zmiennych (np. BMI i Cholesterol). Stwórz wykres punktowy (Wstawianie -> Wykresy -> Wykres punktowy (XY)).
Kliknij prawym przyciskiem myszy na punkty wykresu i wybierz „Dodaj linię trendu". Zanim zaakceptujesz zaznacz „wyświetl równanie na wykresie" oraz „wyświetl wartości R-kwadrat na wykresie".
Zinterpretuj wynik.

b) Regresja liniowa w Analysis ToolPak
Przejdź do zakładki Dane i wybierz Analiza danych->Regresja
Wprowadź: zmienną zależną Y (poziom cholesterolu), zmienną niezależną X (BMI). Zaznacz „zakres wyjściowy" i wybierz miejsce w arkuszu, gdzie mają pojawić się wyniki. Dodatkowo zaznacz „składniki resztowe" oraz „rozkład prawdopodobieństwa normalnego".
Odczytaj współczynniki modelu, R^2, oraz wartość p (p-value) dla zmiennych.
Porównaj otrzymane wyniki z tymi z poprzedniego punktu ćwiczenia. Zauważ, że:

- Wielokrotność R = współczynnik korelacji
- R kwadrat = współczynnik determinacji, który informuje, jaka część wariancji zmiennej zależnej (Y) jest wyjaśniona przez zmienne niezależne (X). Czyli jeśli R^2 wynosi 0,80, oznacza to, że 80% zmienności zmiennej zależnej jest wyjaśnione przez model regresji. Pozostałe 20% to błędy losowe.
- Dopasowany R kwadrat = Skorygowany współczynnik determinacji, który uwzględnia liczbę zmiennych w modelu. Jest użyteczny przy porównywaniu modeli z różną liczbą zmiennych.
- Błąd standardowy = to miara rozrzutu punktów wokół linii regresji. Mówi, jak bardzo prognozy modelu różnią się od rzeczywistych wartości.
- t Stat = wartość statystyki t-Studenta dla korelacji Pearsona (do policzenia istotności)
- Przedziały ufności dla każdego współczynnika regresji = wskazują zakres wartości, w którym z określoną pewnością (zwykle 95%) znajduje się prawdziwy współczynnik w populacji. Jeśli przedział ufności dla współczynnika nie zawiera zera, oznacza to, że współczynnik jest statystycznie istotny.

  - Składniki resztowe (residuals) to różnice między obserwowanymi a przewidywanymi wartościami zmiennej zależnej (Y) w wyniku analizy regresji. Mówią one, jak dobrze model dopasowuje się do danych. Małe reszty wskazują, że model dobrze dopasowuje się do danych.
    Wyniki reszt mogą być używane do:
    - Wykrywania punktów odstających – wartości resztowe, które są znacznie większe niż pozostałe, mogą wskazywać na wartości odstające w danych.
    - Weryfikacji założeń modelu – jeśli reszty mają rozkład normalny, oznacza to, że spełnione są założenia regresji, takie jak normalność błędów.
- Rozkład prawdopodobieństwa normalnego = W kontekście analizy regresji, rozmieszczenie prawdopodobieństwa normalnego odnosi się do rozkładu reszt (błędów) w modelu regresji. W regresji liniowej zakłada się, że wartości resztowe posiadają rozkład normalny. Wykres rozkładu prawdopodobieństwa normalnego pozwala na szybką wizualną ocenę zgodności reszt z rozkładem normalnym (wykres skumulowany). Jeśli nie posiadają one rozkładu normalnego, to nastąpią odstępstwa od linii prostej. Na tym wykresie ujawnią się również obserwacje odstające (nietypowe).

3. Formatowanie warunkowe w Excelu

Policz ciśnienie tętna (różnica między ciśnieniem skurczowym a rozkurczowym) i użyj formatowania warunkowego (Narzędzia główne->Style->Formatowanie warunkowe) żeby wyróżnić niebezpieczne dla zdrowia wartości (czyli takie gdzie ciśnienie tętna < 25 oraz > 60).

4. Obliczenie współczynnika korelacji w języku R

Przygotuj dane wejściowe korzystając z pliku dane_cw11.xlsx oraz na przykład z konwertera tablic online: https://tableconvert.com/excel-to-rdataframe (pamiętaj aby kopiować dane z Excela w czarne pole w konwerterze).

Dane powinny być przygotowane w następującej formie (zwróć uwagę na modyfikację nazw kolumn aby nie zawierały polskich literek oraz spacji):

```
dane <- data.frame(
  Wiek = c("32", "45", "28", "53", "40", "38", "55", "60", "29", "47", "35", "52", "43", "50", "62", "37", "48", "41", "59", "46"),
  BMI = c("22,1", "25,3", "19,8", "30,5", "27,2", "21,5", "34", "29,1", "20,4", "26,2", "23,7", "28,5", "24,8", "31,2", "33,5", "22,9", "27,8", "25", "29,9", "30,3"),
  Cholesterol = c("225", "233", "178", "305", "275", "195", "340", "291", "204", "274", "234", "275", "249", "300", "322", "229", "278", "250", "297", "296"),
  Cisnienie_skurczowe = c("120", "130", "118", "140", "135", "122", "150", "145", "125", "130", "127", "138", "130", "142", "148", "123", "142", "131", "144", "140"),
  Cisnienie_rozkurczowe = c("80", "85", "78", "90", "88", "82", "95", "92", "80", "87", "84", "90", "86", "89", "94", "81", "88", "85", "91", "89")
)
dane$Wiek <- as.numeric(gsub(",", ".", dane$Wiek))
dane$BMI <- as.numeric(gsub(",", ".", dane$BMI))
dane$Cholesterol <- as.numeric(gsub(",", ".", dane$Cholesterol))
dane$Cisnienie_skurczowe <- as.numeric(gsub(",", ".", dane$Cisnienie_skurczowe))
```

```
dane$Cisnienie_rozkurczowe <- as.numeric(gsub(",", ".", dane$Cisnienie_rozkurczowe))
print(dane)
```

Przykładowy kompilator online języka R: https://rdrr.io/snippets/

a) Policz współczynnik korelacji między BMI oraz cholesterolem:
```
cor(dane$BMI, dane$Cholesterol)
```
b) Test istotności korelacji:
```
cor.test(dane$BMI, dane$Cholesterol)
```
c) Wykres punktowy:
```
plot(dane$BMI, dane$Cholesterol, main="Zaleznosc BMI od cholesterolu",
     xlab="BMI", ylab="Poziom cholesterolu", pch=19)
abline(lm(Cholesterol ~ BMI, data=dane), col="blue")
```

5. Budowa modelu regresji liniowej w języku R

d) Zbuduj model regresji:
```
model <- lm(Cholesterol ~ BMI, data=dane)
summary(model)
```
e) Predykcja nowych wartości:
```
nowe_dane <- data.frame(BMI = c(23, 28, 33))
predict(model, newdata=nowe_dane)
```
f) Wykres diagnostyczny:
```
par(mfrow=c(2, 2))
plot(model)
```
Porównaj otrzymane wyniki z tymi uzyskanymi z Excela.

6. Analiza zależności między BMI a ciśnieniem skurczowym

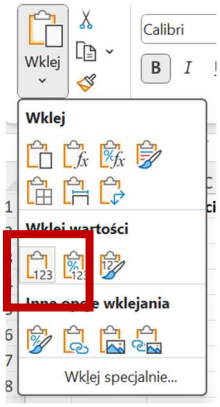
W celu utrwalenia umiejętności wykonaj powyższe analizy dla drugiego przykładu: zależności BMI oraz ciśnienia skurczowego. Skorzystaj z Excela (pakiet Analysis ToolPak) lub języka R.

7. Analiza logistyczna

Pracujesz jako analityk danych medycznych. Otrzymałeś dane od lekarza badającego wpływ BMI oraz wieku pacjentów na ryzyko wystąpienia nadciśnienia tętniczego. Twoim zadaniem jest opracowanie modelu, który oszacuje prawdopodobieństwo wystąpienia nadciśnienia.
Dane wejściowe będą na podstawie pliku dane_cw11.xlsx.

a) Przygotuj tabelę z następującymi zmiennymi:
- Wiek (lata)
- BMI (kg/m²)
- Nadciśnienie: binarna zmienna zależna (0 lub FAŁSZ = brak nadciśnienia, 1 lub PRAWDA = występuje nadciśnienie). Określ, czy pacjent ma nadciśnienie zgodnie z kryteriami, że nadciśnienie występuje gdy: ciśnienie skurczowe ≥ 140 mmHg lub ciśnienie rozkurczowe ≥ 90 mmHg. Formuła do sprawdzenia nadciśnienia powinna wyglądać tak: **=JEŻELI(LUB(D2>=140;E2>=90);1;0)**, gdzie D2 to pole zawierające ciśnienie skurczowe a E2 rozkurczowe.

Przekopiuj te trzy kolumny do nowego arkusza. Uwaga: przekopiuj wartości jak na rysunku obok (zwykłe kopiowanie przekopiuje tylko formuły).



b) Oszacuj prawdopodobieństwo wystąpienia nadciśnienia
W Excelu nie ma wbudowanego narzędzia do regresji logistycznej, ale można ją wykonać za pomocą dodatku Solver.

- Dodaj nową kolumnę Prawdopodobieństwo (np. w kolumnie D), w której oszacujesz prawdopodobieństwo wystąpienia nadciśnienia dla każdego wiersza na podstawie modelu regresji logistycznej.
  Wprowadź formułę:
  =1/(1+EXP(-(H1 + H2*A2 + H3*B2)))
  gdzie:
  - H1 to wyraz wolny,
  - H2 to współczynnik dla Wiek,
  - H3 to współczynnik dla BMI.

Przykładowy wygląd arkusza:

D2 =1/(1+EXP(-(H1 + H2*A2 + H3*B2)))

	A	B	C	D	E	F	G	H
1	Wiek	BMI	Czy nadciśnienie				a	0
2	32	22,1	0	0,5			b	0
3	45	25,3	0	0,5			c	0
4	28	19,8	0	0,5				
5	53	30,5	0	0,5				
6	40	27,2	0	0,5				

c) Dodaj kolumnę Log-Likelihood (np. w kolumnie E), aby obliczyć miarę dopasowania modelu. Wprowadź formułę:
=C2*LN(D2) + (1-C2)*LN(1-D2)
gdzie C2 to zmienna zależna (Nadciśnienie), a D2 to prawdopodobieństwo z modelu.
Policz sumę Log-Likelihood, np w komórce E22:
=SUMA(E2:E21)
Otwórz Solver (Dane > Solver).
Skonfiguruj Solver:
- Maksymalizuj
- Komórka celu**:** suma kolumny Log-Likelihood (np. komórka E22 zawierająca sumę całej kolumny E).
- Przez zmienianie komórek zmiennych**:** komórki H1:H3.
- Ograniczenia: Brak

Kliknij „Rozwiąż” i zapisz wyniki.

d) Za pomocą modelu policz prawdopodobieństwo że osoba, która ma 28 lat i BMI 28,3 ma nadciśnienie. W tym celu wykorzystaj wzór na prawdopodobieństwo (kolumna D) ze znalezionymi przez Solver parametrami a,b,c (np. dostaw na końcu wiersz z podanym wiekiem i BMI).