ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Analiza współzależności

- Współczynnik korelacji liniowej Pearsona
- Współczynnik korelacji rang Spearmana

Analiza zależności

- Liniowa funkcja regresji

Badanie niezależności dwóch cech jakościowych

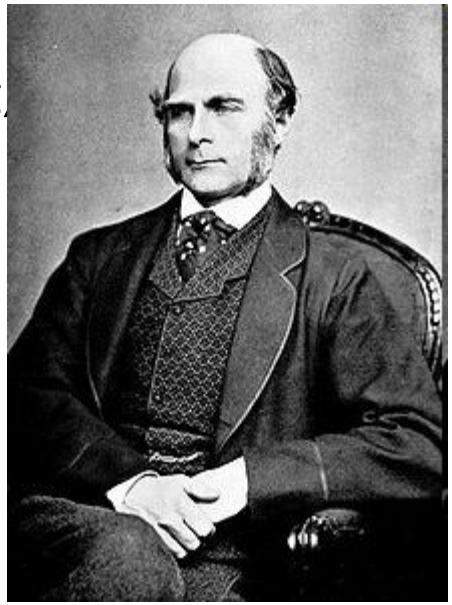
ISTOTA KORELACJI I REGRESJI

- **KORELACJA** daje możliwość stwierdzenia, czy istnieje związek (niekoniecznie przyczynowo-skutkowy) miedzy badanymi cechami (zmiennymi) oraz jaka jest jego siła i kierunek

- **REGRESJA** daje możliwość oszacowania (estymacji) wartości jednej cechy (zmiennej zależnej, objaśnianej) na podstawie wartości przyjmowanych przez drugą cechę (zmienną niezależną, objaśniającą)

- **FUNKCJA REGRESJI,** której parametry można oszacować przy pomocy metody najmniejszych kwadratów (MNK). Równanie opisujące związek statystyczny między zmiennymi nazywa się równaniem lub modelem regresji.

- Sir Francis Galton – 1822-1911, prekursor badań nad inteligencją, statystyk, meteorolog, antropolog, kryminolog. Pisarz, lekarz.
- W 1899 r. w pracy „Naturalna dziedziczność” ogłosił, że rozmiary nasion groszku pachnącego mają tendencję w kolejnych generacjach do powracania (*to regress*) do swego średniego rozmiaru, podobnego związku dopatrzył się także między wzrostem syna i ojca itd.
- Dopasowywał do tych par liczb linię prostą opisującą tę zależność



Zależność przyczynowa – rodzaj zależności, w której jesteśmy w stanie wskazać, która ze zmiennych stanowi przyczynę zmian, a która ilustruje skutek. Przykładem zależności przyczynowej może być związek pomiędzy stażem pracy (przyczyna) i wysokością zarobków (skutek).

Zależność pozorna – pomiędzy dwoma zjawiskami wydaje się istnieć zależność, ale jest ona wywołana istnieniem wspólnej przyczyny. Przykładowo waga i poziom cholesterolu w organizmie wydają się być powiązane ze sobą, niemniej jednak jest to zależność pozorna. W rzeczywistości posiadają wspólną przyczynę – ilość i rodzaj spożywanych produktów

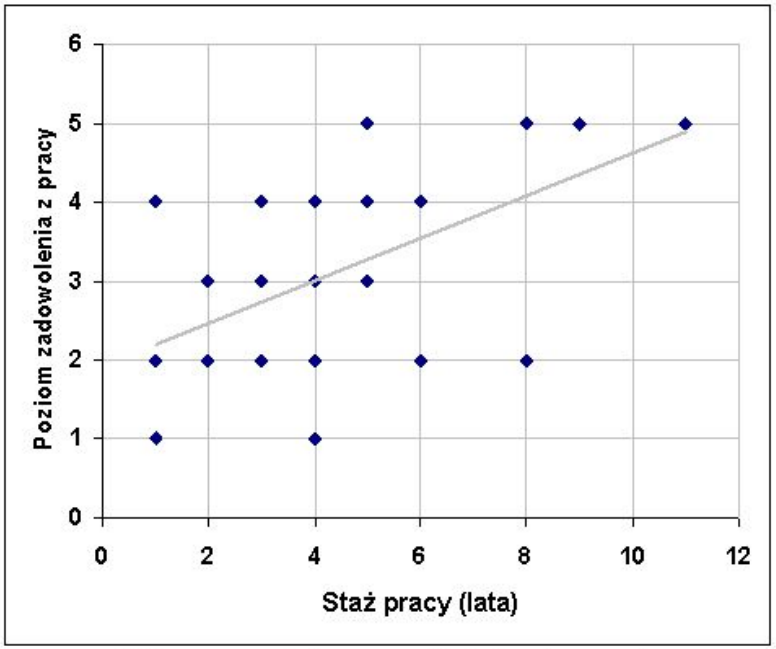
Zależność korelacyjna – zależność w której dla konkretnej wartości jednej zmiennej Xi (zmienna objaśniająca) odpowiada średnia arytmetyczna z kilku wartości drugiej zmiennej Y1, Y2, ...(zmienna objaśniania).

Zmienna niezależna – zmienna która wywołuje zmiany, stanowi ich przyczynę.

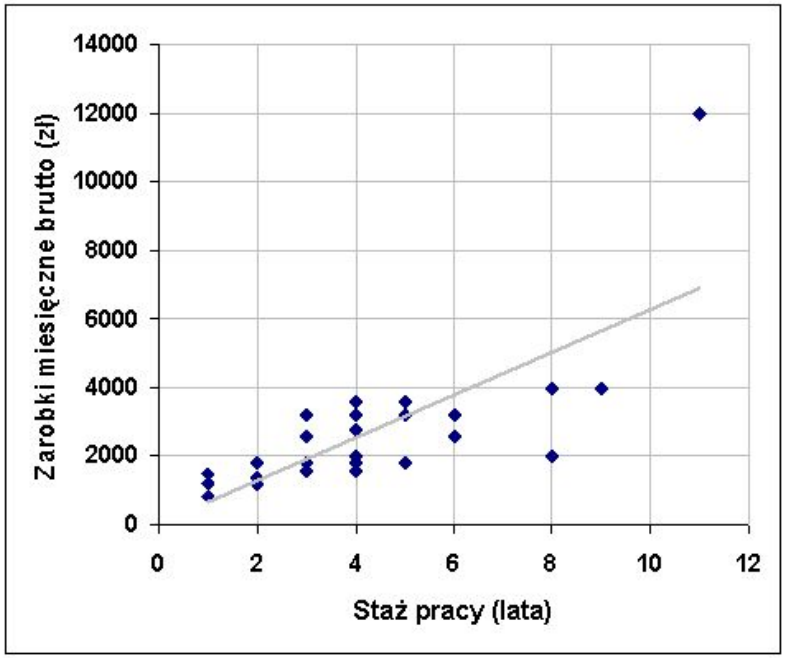
Zmienna zależna – zmienna, której wartości są w mniejszym lub większym stopniu kształtowane przez zmienną niezależną (zmienne niezależne).

Stwierdzenie braku zależności w jednych okolicznościach, nie przesądza o jej nieistnieniu w innych okolicznościach

Wykres korelacyjny (**rozrzutu**) – dla każdego i-tego przypadku nanosimy na układ współrzędnych punkt o współrzędnych (X_i, Y_i), gdzie Xi i Yi to kolejne wartości badanych zmiennych.



Źródło: opracowanie własne na podstawie danych umownych.



Źródło: opracowanie własne na podstawie danych umownych.

WSPÓŁCZYNNIK KORELACJI PEARSONA

Dla zmiennych ilościowych, mierzonych przy pomocy skali przedziałowej lub ilorazowej do określania współzależności najczęściej wykorzystuje się współczynnik korelacji liniowe Pearsona (zakładają, że zależność ma charakter liniowe). Aby obliczyć współczynnik korelacji liniowej, zwykle wcześniej musimy wyznaczyć tak zwaną kowariancję.

Kowariancja – miara współzależności, wyznaczana jako średnia arytmetyczna iloczynu odchyleń wartości zmiennych X i Y od ich średnich arytmetycznych. Kowariancję oznaczamy symbolem **cov(x, y)**.

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Gdzie:

cov(x,y)	- kowariancja dla zmiennych X i Y
n	- liczba obserwacji
x_i, y_i	- wartość i-tej obserwacji dla zmiennych X i Y
$\bar{x}, \bar{y}$	- średnia arytmetyczna dla wartości zmiennych X i Y

Przykład WSPÓŁCZYNNIK KORELACJI PEARSONA

Dla sześciu studentów zmierzono czas pisania egzaminu oraz uzyskaną liczbę punktów. Obliczenia rozpoczynamy od ustalenia średnich dla zmiennej X (czas pisania) oraz Y (liczba punktów):

Kolejne obliczenia wykonujemy korzystając z tabelki:

$\bar{x} = 43,83$

$\bar{y} = 77,17$

Kolejne obliczenia wykonujemy korzystając z tabelki:

Czas pisania egzaminu (x_i)	Liczba punktów (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
51	65	7,17	-12,17	-87,19
35	89	-8,83	11,83	-104,53
48	73	4,17	-4,17	-17,36
39	84	-4,83	6,83	-33,03
45	78	1,17	0,83	0,97
45	74	1,17	-3,17	-3,69
		$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$		**-244,83**

WSPÓŁCZYNNIK KORELACJI PEARSONA

Następnie uzyskany wynik dzielimy przez liczbę obserwacji uzyskując wartość kowariancji:

cov(x, y) = -244,83/6 = -40,81

Wartość kowariancji wskazuje na ujemną zależność korelacyjną pomiędzy czasem pisania egzaminu a liczbą otrzymanych punktów (patrz kolejny slajd).

cov(x,y) = 0	brak jakiejkolwiek zależności pomiędzy badanymi zmiennymi (zmienne nieskorelowane)
cov(x,y) < 0	istnieje ujemna zależność pomiędzy badanymi zmiennymi
cov(x,y) > 0	istnieje dodatnia zależność pomiędzy badanymi zmiennymi

Wartość kowariancji zależy od rzędu wielkości, w jakich wyrażone są obie zmienne, dlatego też nie można jej wykorzystywać bezpośrednio do porównań. Wykorzystać możemy w tym celu współczynnik korelacji liniowej Pearsona.

WSPÓŁCZYNNIK KORELACJI PEARSONA

Współczynnik korelacji liniowej Pearsona – jest miarą siły związku prostoliniowego między dwiema cechami mierzalnymi. Związkiem prostoliniowym nazywamy taką zależność, w której jednostkowym przyrostom jednej zmiennej (zmienna niezależna - X) towarzyszy średnio, stały przyrost drugiej zmiennej (zmienna zależna - Y). Współczynnik korelacji liniowej oznaczamy symbolem $\mathbf{r_{xy}}$. [4]

Współczynnik korelacji liniowej Pearsona liczymy poprzez standaryzację kowariancji, w sposób następujący:

$$r_{xy} = r_{yx} = \frac{cov(x,y)}{s(x) \cdot s(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Gdzie:
s(x), s(y) – odchylenia standardowe dla wartości zmiennej x i y.
Pozostałe oznaczenia jak wcześniej.

WSPÓŁCZYNNIK KORELACJI PEARSONA

Przykład

Bazując na wcześniejszym przykładzie, znamy już średnie arytmetyczne dla zmiennych oraz kowariancję.

cov(x, y) = -40,81

Następnie wyznaczamy odchylenia standardowe dla zmiennych X i Y (proszę wrócić do modułu trzeciego w razie problemów):

s(x) = 5,37
s(y) = 7,78

Współczynnik korelacji liniowej Pearsona obliczamy dzieląc kowariancję przez iloczyn odchyleń standardowych:

$$r_{xy} = -40,81 / (5,37 \times 7,78) = -0,9778$$

Wartość współczynnika korelacji liniowej Pearsona wskazuje na bardzo wyraźną ujemną zależność korelacyjną pomiędzy czasem pisania egzaminu a liczbą otrzymanych punktów (patrz kolejny slajd).

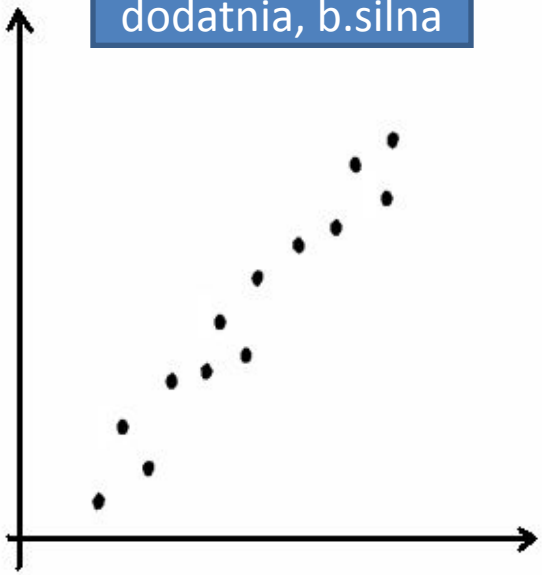
WSPÓŁCZYNNIK KORELACJI PEARSONA

Współczynnik korelacji liniowej Pearsona przyjmuje wartości z przedziału -1 do +1. Jeżeli współczynnik:

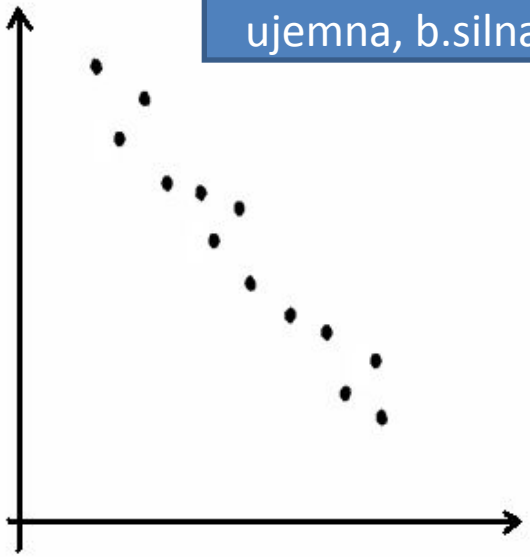
$r_{xy} = 0$	brak jakiejkolwiek zależności pomiędzy badanymi zmiennymi (zmienne nieskorelowane)
$0 < r_{xy} < 0,3$	istnieje niewyraźna dodatnia zależność pomiędzy badanymi zmiennymi
$0,3 < r_{xy} < 0,5$	istnieje średnia dodatnia zależność pomiędzy badanymi zmiennymi
$0,5 < r_{xy} < 1$	istnieje wyraźna dodatnia zależność pomiędzy badanymi zmiennymi
$r_{xy} = 1$	zależność korelacyjna przechodzi w zależność funkcyjną – pomiędzy badanymi zmiennymi istnieje doskonała dodatnia zależność (dla każdej wartości zmiennej x możemy wskazać jednoznacznie wartość dla zmiennej y)

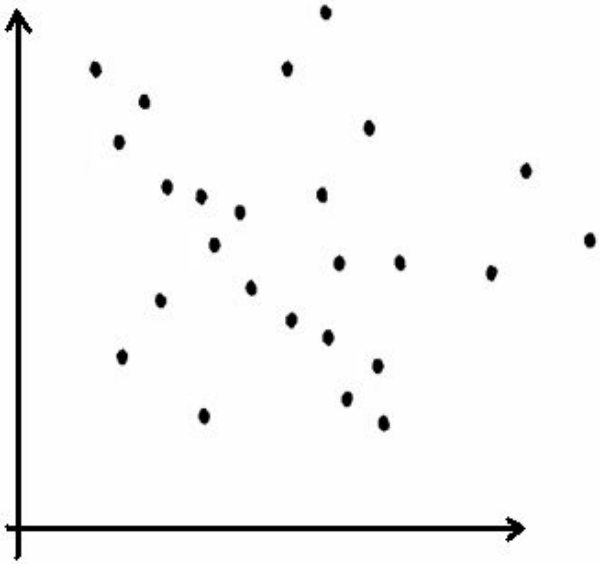
Analogicznie interpretujemy wartości współczynnika mniejsze od zera, z tą różnicą że wówczas mamy doczynienia z ujemną zależnością.

Korelacja dodatnia, b.silna

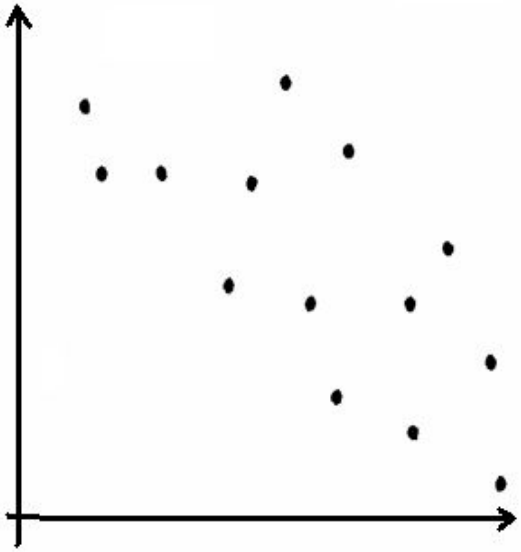


Korelacja ujemna, b.silna





Korelacja niewyraźna, znikoma



Korelacja ujemna, średnia

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

Współczynnik korelacji rang Spearmana służy do opisu siły korelacji dwóch cech w przypadku gdy:

- Cechy są mierzalne, a badana zbiorowość jest nieliczna.
- Cechy mają charakter jakościowy i istnieje możliwość ich uporządkowania.

Współczynnik korelacji rang Spearmana stosuje się do analizy współzależności obiektów pod względem cech X i Y.

Kolejne etapy wyznaczania współczynnika korelacji rang Spearmana są następujące:

1. Jednostki danej zbiorowości statystycznej, ze względu na wielkość odpowiadającej im pierwszej cechy, porządkuje się.
2. Tak uporządkowanym ze względem na pierwszą cechę jednostkom, przypisuje się kolejne numery począwszy od 1. Jeżeli kilka jednostek ma tę samą wielkość cechy, wtedy z odpowiadających im kolejnych rang oblicza się średnią arytmetyczną i przydziela wszystkim jednostkom, z których ta średnia została obliczona. Następna jednostka otrzymuje już najbliższą, niewykorzystaną dotąd rangę. Ostatni numer powinien równać się łącznej liczbie jednostek.
3. Następnie dla jednostek drugiej cechy w analogiczny sposób przypisuje się numery począwszy od 1 (dla jednostki o najniższej lub najwyższej wartości).

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

- Dla tak utworzonych par rang, oblicza się różnicę pomiędzy nimi a następnie podnosi do kwadratu i sumuje wyniki. Otrzymujemy w ten sposób element, który można oznaczyć jako:

$$\sum_{i=1}^{n} d_i^2$$

- A następnie podstawia ten element do wzoru na współczynnik korelacji rang Spearmana:

$$r = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Gdzie:

$d_i = r_{1i} - r_{2i}$ - różnica między parami odpowiadających sobie rang

r_{1i} – ranga i-tego obiektu w pierwszym uporządkowaniu

r_{2i} – ranga i-tego obiektu w drugim uporządkowaniu

n - liczba badanych obiektów

Współczynnik korelacji rang Spearmana przyjmuje wartości z przedziału <-1, 1>. Im bliższy jest on liczbie 1 lub -1, tym silniejsza jest analizowana zależność.

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

Przykład

Dla sześciu studentów zmierzono czas pisania egzaminu oraz uzyskaną liczbę punktów, co przedstawiono w poniższej tabeli:

Student	Czas pisania egzaminu (minut)	Liczba uzyskanych punktów
1	51	65
2	35	89
3	48	73
4	39	84
5	45	78
6	45	74

Wartości przyjmowane przez pierwszą zmienną uporządkowano niemalejąco. Następnie dla obu zmiennych (czas i liczba punktów) przypisano rangi (od najmniejszej do największej wartości), obliczono różnicę pomiędzy rangami, podniesiono tak obliczoną różnicę od kwadratu i wyniki zsumowano.

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

Warto zauważyć, że czas pisania dla studentów 5 i 6 jest taki sam i wynosi 45 minut. Jest to trzeci i czwarty czas pisania egzaminu, w związku z czym przypisujemy jednakowe rangi stanowiące średnią arytmetyczną wartości 3 i 4.

Student	Czas pisania egzaminu (ranga r_{1i})	Liczba uzyskanych punktów (ranga r_{2i})	Róznica rang ($d_i = r_{1i} - r_{2i}$)	Różnica podniesiona do kwadratu (d_i^2)
2	1	6	-5	25
4	2	5	-3	9
6	3,5	3	0,5	0,25
5	3,5	4	-0,5	0,25
3	5	2	3	9
1	6	1	5	25
			Suma	68,5

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

Obliczoną sumę kwadratów różnic podstawiamy do wzoru:

$$r = 1 - \frac{6 \cdot 68{,}5}{6(6^2 - 1)} = -0{,}96$$

Współczynnik równy -0,96 świadczy o istnieniu wyraźnej negatywnej zależności pomiędzy czasem pisania egzaminu a ilością punktów. Im student dłużej pisze, tym statystycznie rzecz biorąc otrzymuje mniej punktów (co można wytłumaczyć faktem, że najlepiej przygotowani studenci kończą pisać egzamin wcześniej).

FUNKCJA REGRESJI

W modelach regresji zależność pomiędzy jedną lub większą ilością zmiennych niezależnych (predykatory, zmienne wyjaśniające) a zmienną zależną (zmienna wyjaśniana) przedstawiamy w postaci tak zwanej funkcji regresji.

Poniżej przedstawiono przykłady wykorzystania modeli regresji do rozwiązywania praktycznych problemów:

Określenie zależności pomiędzy wiekiem, poziomem wykształcenia (mierzonym na przykład przez liczbę lat), stażem pracy a wysokością zarobków w danej branży.
Określeniem wpływu działań marketingowych (mierzonych na przykład wydatkami na reklamy telewizyjne, prasowe, billboardy, etc.) na przyszłą sprzedaż produktu.
Określenie wpływu wieku, wagi, aktywności ruchowej (mierzonej na przykład liczbą godzin w tygodniu przeznaczoną na uprawianie sportu) a kondycją fizyczną (mierzoną na przykład wynikiem biegu na 1km).

K*arol Fryderyk Gauss*, ur. w 1777 roku w Niemczech. Ojciec Karola był pomocnikiem murarskim i swojego syna początkowo przeznaczał do podobnej kariery. Na szczęście niepospolity talent młodziutkiego Gaussa objawił się na tyle wcześnie i w sposób tak ewidentny, że znalazł się oświecony i możny sponsor, dzięki któremu matematyka nie straciła jednego ze swoich najwybitniejszych uczonych. Nauczycielu matematyki kazał swoim uczniom (8-9letnim) obliczyć sumę liczb od 1 do 100. Karol po pięciu minutach przedstawił kartkę z rzeczywiście króciutkim wywodem:

1	2	3	…	50
100	99	98	…	51
101	101	101	…	101

101x50=5050

Jeszcze jako uczeń gimnazjum Gauss sformułował metodę najmniejszych kwadratów

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left[y_i-\left(a\cdot x_i+b\right)\right]^2=min$$

FUNKCJA REGRESJI

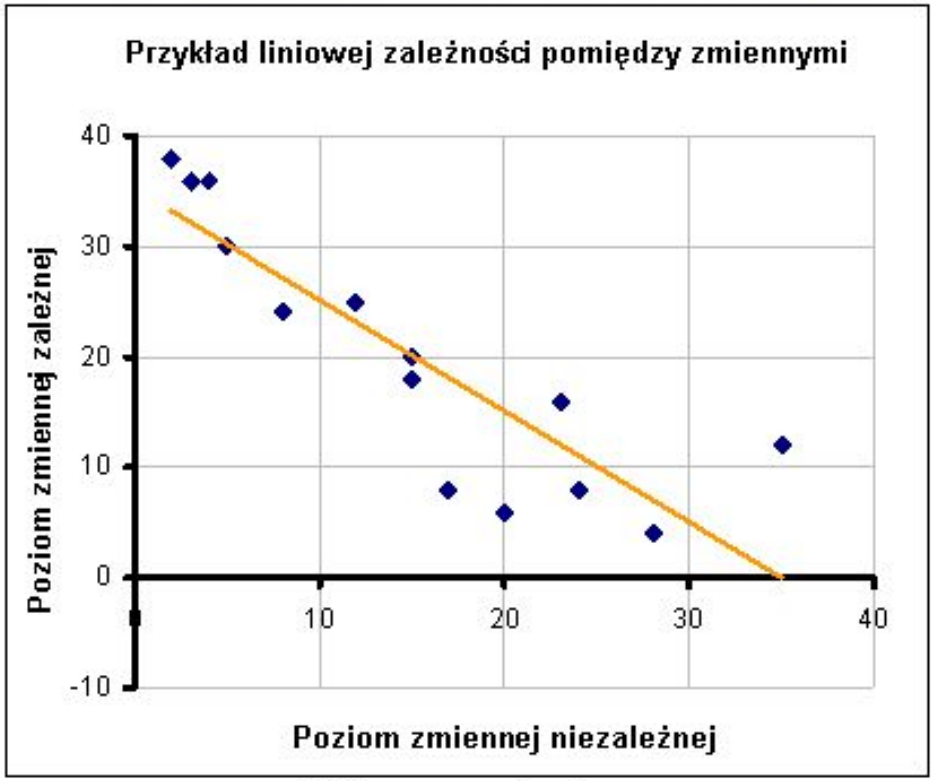
Funkcja regresji - to narzędzie do badania powiązań między zmiennymi. Funkcja regresji to analityczny wyraz przyporządkowania średnich wartości zmiennej zależnej konkretnym wartością zmiennej niezależnej.

Dużym problemem jest wybór postaci analitycznej funkcji dla danego problemu. Ułatwieniem może być sporządzenie m.in. wykresu rozrzutu, gdzie dla każdej (i-tej) pary wartości zmiennej niezależnej (X) i zmiennej zależnej (Y) tworzymy punkt o współrzędnych Xi, Yi.

Jeżeli zmiennych niezależnych jest więcej, wówczas konstruujemy odpowiednio większą ilość wykresów rozrzutu, przedstawiających zależność pomiędzy każdą zmienną niezależną (oś pozioma) a zmienną niezależną. Z wykresu (wykresów) odczytujemy prawdopodobny rodzaj zależności pomiędzy zmiennymi niezależnymi a zmienną zależną.

FUNKCJA REGRESJI

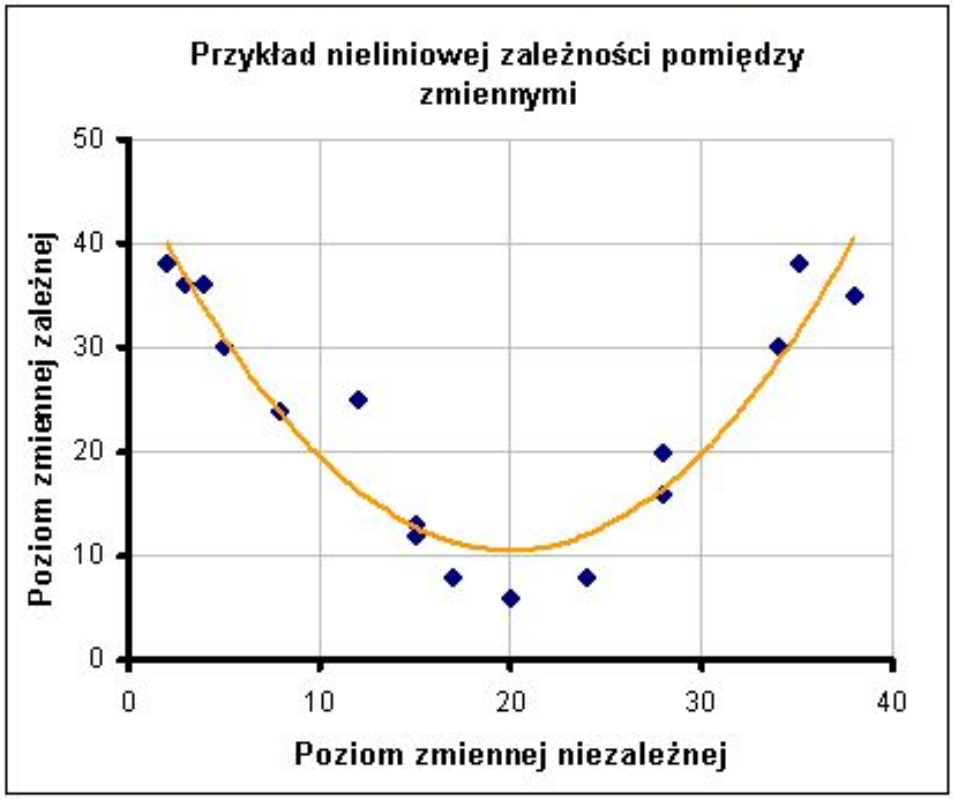
Jeżeli chmura punktów układa się w przybliżeniu wzdłuż linii prostej (co zostało pokazane na poniższym wykresie), wówczas możemy wykorzystać liniową funkcję regresji.



Źródło: opracowanie własne.

FUNKCJA REGRESJI

Jeżeli chmura punktów nie układa się wzdłuż linii prostej, wówczas należy wykorzystać inną analityczną postać funkcji regresji (na przykład funkcje potęgowe, logarytmiczne, wielomianowe czy też wykładnicze).



Źródło: opracowanie własne.

FUNKCJA REGRESJI

Mamy do czynienia tylko z jedną zmienną niezależną X.
Zależność pomiędzy zmienną niezależną X a zmienną zależną Y ma charakter liniowy.

Naszym zadaniem jest wyznaczenie liniowej funkcji regresji, o ogólnej postaci:

y = a + bx

Gdzie:

y - wartość przewidywana na podstawie wartości **x**
a - parametr a jest nazywany wyrazem wolnym i odpowiada wartości funkcji y dla argumentu x = 0
b - współczynnik kierunkowy, który decyduje o tym, czy funkcja jest rosnąca, czy malejąca oraz jak szybko następują zmiany (jeśli b jest dodatnie, to funkcja jest rosnąca – to znaczy, im większe wartości zmiennej x, tym większe wartości funkcji, czyli y)

Do wyznaczenia parametrów tej funkcji (a i b) wykorzystuje się metodę najmniejszych kwadratów.

FUNKCJA REGRESJI

Współczynniki regresji liniowej szacuje się za pomocą metody najmniejszych kwadratów w ten sposób, aby suma kwadratów odchyleń wartości teoretycznych (wyznaczonych w oparciu o funkcję regresji) i zaobserwowanych jest najmniejsza. Parametry a i b dla liniowej funkcji regresji możemy obliczyć korzystając z następujących wzorów:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

Gdzie:

n	- liczba obserwacji
x_i, y_i	- wartość i-tej obserwacji dla zmiennych X i Y
$\bar{x}, \bar{y}$	- średnia arytmetyczna dla wartości zmiennych X i Y

Pomiędzy parametrami funkcji regresji liniowej, a wcześniej omówionym współczynnikiem korelacji liniowej Pearsona istnieje ścisła zależność (dlatego nazywamy go liniowym):

$$b = r_{xy}\frac{s_y}{s_x} \qquad \text{oraz} \qquad a = r_{xy}\frac{s_x}{s_y}$$

FUNKCJA REGRESJI

Po wyznaczeniu parametrów funkcji regresji liniowej należy ocenić **poziom dopasowania** funkcji regresji do rzeczywistych danych. Sprowadza się to do odniesienia generowanych przez funkcję regresji **wartości teoretycznych** do **wartości zaobserwowanych**. Wykorzystuje się w tym celu szereg miar, do najczęściej stosowanych należą: odchylenie standardowe reszt, współczynnik zbieżności oraz współczynnik determinacji.

Wartości teoretyczne obliczamy podstawiając do funkcji regresji liniowej wartości zmiennej niezależnej X.

Przykład

Dla pewnej funkcji regresji liniowej:

$$y = 250 - 2x$$

Obliczamy wartości teoretyczne dla zmiennej niezależnej **x** równej 10 oraz 11.

Dla x = 10 otrzymujemy: $y = 250 - 2*10 = 230$
Dla x = 11 otrzymujemy: $y = 250 - 2*11 = 228$

W wyjaśnianiu wielu zjawisk istotną rolę odgrywają zmienne niemierzalne, tj. *jakościowe. I tak, na wielkość popytu na dany produkt oprócz jego walorów użytkowych* wielki wpływ ma *marka. Dotyczy to w szczególności takich produktów jak samochody,* odzież, zegarki czy sprzęt elektroniczny. W analizie wydajności pracy w rozmaitych zawodach istotną rolę odgrywa *płeć pracownika. Ma ona także wpływ na wynagrodzenie.*
To ostanie z kolei w sposób oczywisty zależy od *stanowiska. Wielkość dochodów* ludności zależy od *kraju, który ona zamieszkuje, itd.*
Podobne wielkości występują przy analizie rozmaitych procesów chemicznych czy fizycznych (np. rodzaj użytego tworzywa, sposób (technika) obróbki, itp.)
Wartości zmiennej jakościowej nazywamy *kategoriami lub wariantami.*
Jeśli różnych kategorii zmiennej jakościowej jest stosunkowo niewiele, to zmienną taką możemy łatwo włączyć do modelu regresji.

WSPÓŁZALEŻNOŚC CECH JAKOŚCIOWYCH

Dla danych jakościowych, mierzonych na skali nominalnej lub porządkowej analizę współzależności zwykle rozpoczynamy od utworzenia tabeli krzyżowej. W pierwszej kolumnie warianty cechy X, natomiast w pierwszym wierszu tabeli umieszczamy warianty zmiennej Y. Możliwe jest także utworzenie tabeli krzyżowej dla zmiennych ilościowych, mierzonych na skali przedziałowej lub ilorazowej. Wówczas gdy liczba wszystkich przyjmowanych wartości przez zmienną X i Y (liczbę możliwych wartości będziemy oznaczać symbolami k i l) jest względnie mała, wpisujemy je wszystkie w odpowiednie wiersze i kolumny. W przypadku dużej liczby możliwych wartości niezbędne jest ich pogrupowanie przy użyciu przedziałów klasowych.

WSPÓŁZALEŻNOŚC CECH JAKOŚCIOWYCH

W tym przypadku jako zmienną X przyjęliśmy Płeć, natomiast jako zmienną Y przyjęliśmy Ukończenie studiów MBA. Obie zmienne są jakościowe, wyrażane przy pomocy skali nominalnej. Obie posiadają dwa możliwe warianty (k = l = 2).

Płeć	Ukończone studia MBA		Suma końcowa
	Nie	Tak	
Kobieta	13	3	**16**
Mężczyzna	7	7	**14**
Suma końcowa	**20**	**10**	**30**

Źródło: opracowanie własne na podstawie danych umownych.

Z powyższej tabeli możemy odczytać, że spośród wszystkich pracowników firmy objętych badaniem (n=30), 16 stanowią kobiety, natomiast 14 mężczyźni. Ponadto jesteśmy w stanie stwierdzić, że 20 osób nie ukończyło studiów MBA, przy 10 osobach które posiadają tytuł MBA. Komórki położone w środku tabeli (nie na jej brzegach) oznaczają sytuację, w których współwystępują określone warianty cechy X i Y. Na przykład wartość 13 oznacza tyle kobiet, które nie ukończyły studiów MBA.

Z tabeli możemy także odczytać, że połowa mężczyzn pracujących w firmie ukończyła te studia, natomiast w przypadku kobiet są to zaledwie 3 osoby z łącznej liczby 16.