Projekt realizowany
z Narodowym Bankiem Polskim
w ramach programu edukacji ekonomicznej

SIGMA KWADRAT

LUBELSKI KONKURS STATYSTYCZNO-DEMOGRAFICZNY

Projekt realizowany
z Narodowym Bankiem Polskim
w ramach programu edukacji ekonomicznej

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

W rzeczywistości społeczno-gospodarczej obserwuje się często, że określone zjawisko jest przyczynowo uwarunkowane innymi zjawiskami. Inaczej mówiąc zmiany zachodzące w jednym zjawisku wywoływane są zmianami w innych zjawiskach. Ograniczając się jedynie do związków między dwoma zjawiskami jako przykłady można podać takie związki, jak:

- **związek między dochodami a popytem,**
- **związek między dochodami a wydatkami.**
- **związek między stażem pracy a poziomem płacy,**
- **związek między wydajnością pracy a wielkością płac,**
- **związek między zatrudnieniem a wielkością produkcji,**
- **związek między wielkością produkcji a kosztami,**
- **związek między ceną a popytem.**

Związki zachodzące między zjawiskami mogą mieć charakter:

- **funkcyjny,**
- **stochastyczny,**
- **korelacyjny.**

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Związek funkcyjny ma miejsce wówczas, gdy określonej wartości jednej wielkości (cechy) odpowiada określona wartość drugiej wielkości (cechy). Przykładem takiego związku może być zależność należności za zakupiony towar od ilości zakupu.

Związek stochastyczny występuje wówczas, gdy konkretnej wartości jednej wielkości (cechy) odpowiadają różne wartości drugiej wielkości (cechy). Jako przykład takiego związku można traktować sytuację, gdy pracownicy o takim samym stażu pracy otrzymują różne płace. Można więc powiedzieć, że ze związkiem stochastycznym mamy do czynienia, gdy określonej wartości jednej wielkości odpowiada rozkład wartości drugiej wielkości.

Związek korelacyjny występuje wówczas, gdy określonej wartości jednej wielkości (cechy) przyporządkowana jest przeciętna wartość drugiej wielkości (cechy). Na przykład o związku korelacyjnym mówimy, gdy stwierdzamy: pracownicy, którzy mają 10-letni staż pracy, średnio zarabiaj 3200 zł miesięcznie.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Rodzaje zależności:

***funkcyjna** - danej wartości zmiennej niezależnej odpowiada jedna i tylko jedna wartość zmiennej zależnej,*

***stochastyczna** - danej wartości zmiennej niezależnej odpowiada rozkład wartości zmiennej zależnej (inaczej można powiedzieć, że zależność stochastyczna ma miejsce wówczas gdy wartość zmiennej niezależnej wpływa na prawdopodobieństwo zajścia zmiennej zależnej),*

***korelacyjna (statystyczna)** - danej wartości zmiennej niezależnej odpowiada przeciętna wartość zmiennej zależnej.*

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Stwierdzanie istnienia (lub braku) związku korelacyjnego między dwoma zjawiskami można dokonać na podstawie:

- ✓ **porównania kształtowania się szeregów statystycznych szczegółowych charakteryzujących rozpatrywane zjawiska,**
- ✓ **wykresu (metoda graficzna),**
- ✓ **zbudowanej tablicy korelacyjnej, gdy dane o dwóch zjawiskach podane są w formie szeregów rozdzielczych.**

Rodzaje zależności korelacyjnej:

1. ***ze względu na liczbę zmiennych:***
   - **zwykła (prosta, dwuwymiarowa)**
   - **wieloraka (wielokrotna, wielowymiarowa)**
2. ***ze względu na kierunek oddziaływania:***
   - **dodatnia**
   - **ujemna**
3. ***ze względu na kształt powiązań:***
   - **liniowa**
   - **nieliniowa (sprowadzalna do liniowej i całkowicie nieliniowa)**

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

PROSTE METODY STWIERDZANIA ZWIĄZKÓW KORELACYJNYCH

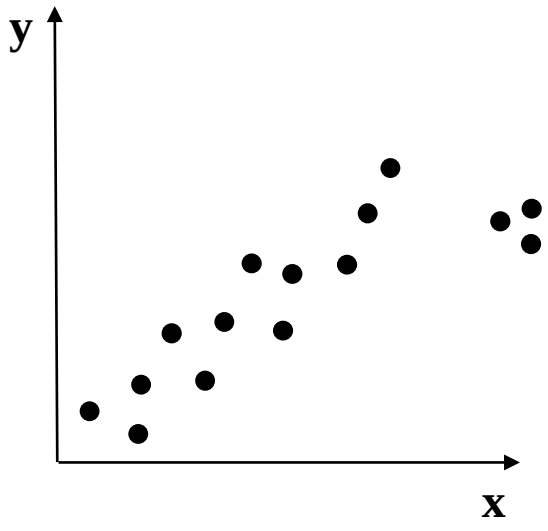
1. **porównanie wzajemnego kształtowania się zmiennych (gdy zmienne są podane w formie szeregów szczegółowych),**
2. **wykres rozrzutu punktów empirycznych,**
3. **tablica korelacyjna (gdy zmienne podane są w formie szeregów rozdzielczych).**

Jeżeli, mając do dyspozycji dwa szeregi szczegółowe stwierdza się, że rosnącym (malejącym) wartościom jednej cechy towarzyszy wzrost (spadek) wartości drugiej cechy, to korelacja pomiędzy tymi cechami jest *dodatnia.*

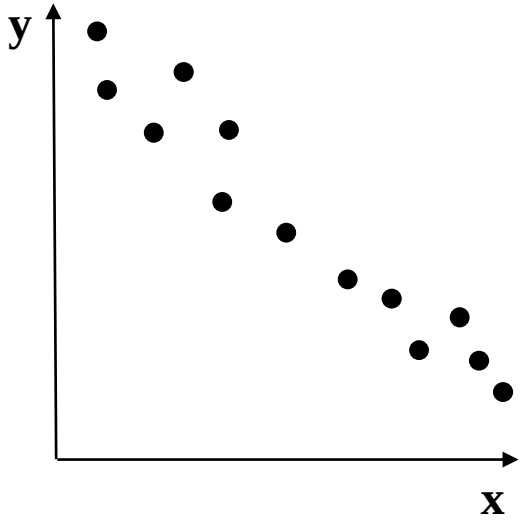
Jeżeli natomiast rosnącym (malejącym) wartościom jednej cechy towarzyszy spadek (wzrost) wartości drugiej cechy, to korelacja jest *ujemna*.

Poniższe wykresy przedstawiają kilka różnych sytuacji, z których wynika występowanie korelacji lub jej brak w przypadku gdy rozpatruje się związek między dwoma zjawiskami.

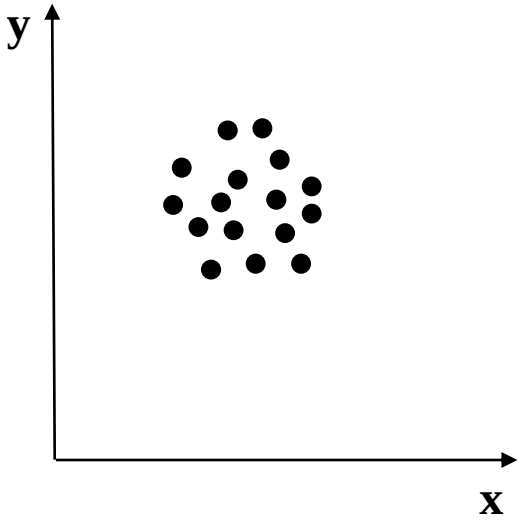
ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK



Związek korelacyjny dodatni



Związek korelacyjny ujemny



Brak korelacji

x – zmienna niezależna (zmienna objaśniająca), /na osi odciętych/

y – zmienna zależna (zmienna objaśniana). /na osi rzędnych/

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

1. **współczynnik korelacji liniowej Persona,**
2. **współczynnik korelacji rang Spearmana,**
3. **współczynnik korelacji wielorakiej,**
4. **inne.**

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji liniowej Pearsona **mierzy związek między zmiennymi x i y, gdy są one mierzalne i podane w postaci szeregów szczegółowych. Opisywany jest wzorem:**

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

Uwaga!!! Współczynnik korelacji Pearsona wskazuje kierunek korelacji, r_{xy} jest wielkością unormowaną i może przyjmować wartości w granicach od -1 do +1. Znak „+" oznacza korelację dodatnią, znak „-" oznacza korelację ujemną. Gdy $r_{xy} = 0$ związek korelacyjny między cechami nie występuje. Charakterystyczną cechą tego współczynnika jest również jego symetryczność tzn. $r_{xy} = r_{yx}$.

W niektórych publikacjach przyjmuje się, że:

$0{,}2 \leq |r| < 0{,}4$, korelacja niska,
$0{,}4 \leq |r| < 0{,}7$, korelacja wyraźna (umiarkowana),
$0{,}7 \leq |r| < 0{,}9$, korelacja silna (znacząca),
$0{,}9 \leq |r| < 1{,}0$, korelacja bardzo silna.

Uwaga!!! Wartość współczynnika bliska „0" nie zawsze oznacza brak zależności, a jedynie brak zależności liniowej.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji liniowej Pearsona c.d.

Kwadrat współczynnika korelacji zwany jest współczynnikiem determinacji i określa, w jakim stopniu zmiany jednej cechy są wyjaśniane przez zmiany drugiej cechy, np. r^2_{xy} = 0,85 oznacza, że w 85,0% zmiany zmiennej y są wyjaśniane zmiennością zmiennej x.

ustalenie na podstawie danych, czy badane zjawiska są zależne

x_i	y_i
2	15
3	16
4	14
5	17
6	20
7	19
8	21
9	23
10	24
11	24

x_i	y_i
2	24
3	24
4	23
5	21
6	19
7	20
8	17
9	14
10	16
11	15

x_i	y_i
2	10
3	10
4	10
5	11
6	10
7	10
8	9
9	10
10	10
11	9

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji liniowej Pearsona c.d.

Wstępnej oceny istnienia lub braku zależności możemy dokonać na podstawie danych empirycznych, przedstawionych w szeregach korelacyjnych. Jeżeli wzrost wartości jednej z cech wywołuje rosnącą tendencję wartości cechy drugiej to cechy te *są zależne*, a kierunek zależności *dodatni*. Jeżeli wzrostowi wartości jednej z cech towarzyszy tendencja malejąca wartości cechy drugiej to stwierdzamy zależność o kierunku *ujemnym*. Jeżeli zmianom wartości jednej z cech nie towarzyszy żadna określona tendencja zmian wartości cechy drugiej, to stwierdzamy *brak zależności.*

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji rang Spearmana

Współczynnik ten stosuje się do opisu siły korelacji dwóch cech, gdy:

1. **cechy są mierzalne, a liczebność badanej zbiorowości jest nieliczna,**
2. **cechy maja charakter jakościowy i istnieje możliwość ich zrangowania, czyli uporządkowania od najmniejszej do największej lub odwrotnie.**

Współczynnik korelacji rang wyznacza się z wzoru:

$$R_{yx} = r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

gdzie: **d_i - różnica między rangami nadanymi wartościom zmiennej x uporządkowanym rosnąco i rangami nadanymi odpowiadającym im wartościom zmiennej y.**
n – liczba obserwacji.

$$-1 \le R_{yx} \le 1$$

$|R_{yx}| = 1$ *to zależność pełna (funkcyjna) wskazuje zarówno siłę jak i kierunek zależności*

$R_{yx} = 0$ *to brak zależności*

Znak określa kierunek:
„+” oznacza, że uporządkowanie wg obu cech jest jednokierunkowe
„-” oznacza, że uporządkowanie jednej cechy jest przeciwne do uporządkowania wg cechy drugiej.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji rang Spearmana

Może on być obliczany zarówno dla cech mierzalnych, jak i niemierzalnych. Warunkiem koniecznym jest możliwość uporządkowania danych rosnąco lub malejąco. Kolejnym wartościom każdej z cech przypisujemy odpowiednią rangę, która wskazuje pozycję danej jednostki w szeregu korelacyjnym.

Własności r_s są takie same jak r_{xy} z tym, że *gdy chcemy go zastosować do badania siły związku między cechami mierzalnymi, musi być spełniony wymóg liniowości związku*.

Należy zwrócić uwagę na fakt, że do wyznaczenia współczynnika korelacji rang nie są potrzebne dane wyjściowe dotyczące rzeczywistych wartości badanych cech, lecz dane (rangi) jakie nadano poszczególnym wartościom rozpatrywanych cech wynikające z ich uporządkowania.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji rang Spearmana

Zadanie

Przeprowadzono badanie zależności wydajności pracy od stopnia zdyscyplinowania pracowników. Otrzymano następujące uszeregowanie wg obu cech:

L.p.	**Zdyscyplinowanie pracowników** x_i	**Wydajność** y_i	dx_i	dy_i
1	A	E	1	5
2	D	D	4	4
3	E	F	5	6
4	B	J	2	10
5	G	K	7	11
6	I	L	9	12
7	C	C	3	3
8	F	I	6	9
9	H	B	8	2
10	J	H	10	8
11	L	A	12	1
12	K	G	11	7

Określić siłę
i kierunek zależności pomiędzy badanymi cechami.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji rang Spearmana

Rozwiązanie: Nadajemy rangi, gdzie A=1, B=2, C=3 itd.

dx_i	dy_i	d_i	d_i^2
1	5	-4	16
4	4	0	0
5	6	-1	1
2	10	-8	64
7	11	-4	16
9	12	-3	9
3	3	0	0
6	9	-3	9
8	2	6	36
10	8	2	4
12	1	11	121
11	7	4	16
-------	----------	----------	**292**

$$R_{yx} = r_s = 1 - \frac{6*292}{12(144-1)} = -0{,}02$$

Wnioski:

Zależność pomiędzy wydajnością pracy, a zdyscyplinowaniem prawie nie istnieje. Kierunek uporządkowania obu cech jest przeciwny.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

MIERNIKI NATĘŻENIA (SIŁY) ZWIĄZKU KORELACYJNEGO

Współczynnik korelacji wielorakiej

Współczynnik korelacji rang wyraża się wzorem: $$r_w = \sqrt{1 - \frac{\det W}{\det R}}$$

Podstawą dla konstrukcji współczynnika korelacji wielorakiej jest macierz W złożona ze współczynników korelacji prostej (pomiędzy zmienną *y* jako zmienną zależną i liczbą *k* zmiennych niezależnych x_{ij}):

$$W = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_k} \\ r_{x_1y} & 1 & r_{x_1x_2} & \dots & r_{x_1x_k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{x_ky} & r_{x_kx_1} & r_{x_kx_2} & \dots & 1 \end{bmatrix} \quad R = \begin{bmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_1} \\ & & 1 & \\ r_{x_kx_1} & r_{x_kx_2} & r_{x_2x_k} & 1 \end{bmatrix}$$

!!! r_w jest liczbą z przedziału <0;1>; informuje o sile związku między zmienną zależną Y a całym zbiorem zmiennych niezależnych $X_1, \dots, X_k$.

gdzie: **r_w - współczynnik korelacji wielorakiej,**

det W - wyznacznik macierzy W,

det R - wyznacznik macierzy R.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

TABLICA KORELACYJNA

y_j / x_i	y_1	y_2	...	y_l	$n_{i\cdot}$	$\bar{y}_i$	$s^2_{i(y)}$
x_1	n_{11}	n_{12}	...	n_{1l}	$n_{1\cdot}$	$\bar{y}_1$	$s^2_{1(y)}$
X_2	N_{21}	N_{22}	...	n_{2l}	$n_{2\cdot}$	$\bar{y}_2$	$s^2_{2(y)}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kl}	$n_{k\cdot}$	$\bar{y}_k$	$s^2_{k(y)}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot l}$	n		
$\bar{x}_j$	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_l$			
$s^2_{j(x)}$	$s^2_{1(x)}$	$s^2_{2(x)}$	...	$s^2_{l(x)}$			

i = 1, 2, ..., k - liczba wierszy
j = 1, 2, ..., l - liczba kolumn

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

TABLICA KORELACYJNA

Charakterystyki rozkładów brzegowych:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i. \qquad S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j n_{.j} \qquad S_y^2 = \frac{1}{n}\sum_{j=1}^{l}(y_j - \bar{y})^2 n_{.j}$$

Charakterystyki rozkładów warunkowych

$$\bar{x}_j = \frac{1}{n_{.j}}\sum_{i=1}^{k} x_i n_{ij} \qquad S_{j(x)}^2 = \frac{1}{n_{.j}}\sum_{i=1}^{k}(x_i - \bar{x}_j)^2 n_{ij}$$

$$\bar{y}_i = \frac{1}{n_{i.}}\sum_{j=1}^{l} y_j n_{ij} \qquad S_{i(y)}^2 = \frac{1}{n_{i.}}\sum_{j=1}^{l}(y_j - \bar{y}_i)^2 n_{ij}$$

Jeżeli charakterystyki rozkładów brzegowych nie równają się charakterystykom rozkładów warunkowych mamy do czynienia z zależnością stochastyczną, jak i korelacyjną.

Uwaga!!! Brak zależności korelacyjnej nie oznacza braku zależności stochastycznej.

Jeżeli $\bar{x} = \bar{x}_j$, $\bar{y} = \bar{y}_i$, $S_y^2 = S_i^2$, $S_x^2 = S_j^2$,

to mamy do czynienia ze stochastyczną niezależnością, a więc i brakiem skorelowania.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Istotność niezależności stochastycznej bada się przy pomocy *nieparametrycznego testu niezależności χ2.*

$H_0 : E(n_{ij} - \hat{n}_{ij}) = 0$ **oznacza brak zależności stochastycznej**

$H_1 : E(n_{ij} - \hat{n}_{ij}) \neq 0$

gdzie: $\hat{n}_{ij} = \frac{n_i n_{\cdot j}}{n}$

Funkcja testowa ma postać: $$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Statystyka χ2 ma rozkład χ2 określony liczbą stopni swobody s = (k - 1)(l - 1).

Jeżeli w procesie weryfikacyjnym odrzucamy H_0, oznacza to, że zmienne są istotnie stochastycznie zależne czyli mogą być skorelowane.

Jeżeli nie precyzuje się, która ze zmiennych pełni rolę zmiennej zależnej, do zmierzenia siły tego związku można wykorzystać współczynnik V- Cramera o postaci:

$$V = \sqrt{\frac{\chi^2}{n(g-1)}}$$ **gdzie: g = min (k,l)**

Współczynnik V przyjmuje wartości z przedziału <0;1> i jeżeli V = 0, to istnieje niezależność zmiennych, V = 1, to zależność jest funkcyjna.

Współczynnik V jest wygodnym miernikiem siły związku,

gdy zmienne są jakościowe.

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Jeżeli ustalamy, że zmienna *y* jest zależna lub zmienna *x* jest zależna i są one mierzalne, to w roli miernika natężenia korelacji między nimi występuje współczynnik korelacji Pearsona z tablicy korelacyjnej, zwany *stosunkiem korelacyjnym* (e_{xy} lub e_{yx}).

Podstawą konstrukcji wzoru na stosunek korelacyjny jest równość wariancyjna, która w przypadku tablicy korelacyjnej ma postać następującą:

$$S_x^2 = \bar{S}_j^2 + S_{\bar{x}_j}^2 = \frac{1}{n}\sum_{j=1}^{l} S_j^2 n_{.j} + \frac{1}{n}\sum_{j=1}^{l} (\bar{x}_j - \bar{x})^2 n_{.j}$$

$$S_y^2 = \bar{S}_i^2 + S_{\bar{y}_i}^2 = \frac{1}{n}\sum_{i=1}^{k} S_i^2 n_{i.} + \frac{1}{n}\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_{i.}$$

Pierwszy składnik równości wariancyjnej to *wariancja wewnątrzgrupowa*, drugi zaś to *wariancja międzygrupowa.*

Wykorzystując elementy równości wariancyjnej mamy:

$$S_x^2 = \bar{S}_j^2 + S_{\bar{x}_j}^2 \;/\, S_x^2 \qquad 1 = \frac{\bar{S}_j^2}{S_x^2} + \frac{S_{\bar{x}_j}^2}{S_x^2}$$

stąd: $$e_{xy} = \sqrt{1 - \frac{\bar{S}_j^2}{S_x^2}} = \frac{S_{\bar{x}_j}}{S_x}$$

Analogicznie: $$e_{yx} = \sqrt{1 - \frac{\bar{S}_i^2}{S_y^2}} = \frac{S_{\bar{y}_i}}{S_y}$$

ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Współczynniki e_{xy} i e_{yx} przyjmują wartości z przedziału; <0;1>, nie wskazują zatem kierunku korelacji. Można je stosować zarówno do korelacji liniowej, jak i nieliniowej (są efektywne).

UWAGA!!! Jeżeli rozkład liczebności w tablicy korelacyjnej wskazuje na liniowość związku pomiędzy zmiennymi, to do analizy zależności można bezpośrednio stosować współczynnik korelacji liniowej Pearsona z tablicy korelacyjnej, który ma postać:

$$r_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{nS_xS_y}$$

Relacja pomiędzy e_{xy} i e_{yx} , a r_{xy} przedstawia się następująco: $r^2_{xy} \leq e^2_{yx}$ **i** $r^2_{xy} \leq e^2_{xy}$

Stąd mamy *współczynnik krzywoliniowości* o postaci:

$$m(yx) = e^2_{yx} - r^2_{xy}$$
$$m(xy) = e^2_{xy} - r^2_{xy}$$

ANALIZA REGRESJI

***Regresja teoretyczna*: równanie matematyczne (konstrukcja formalna, model), przedstawiające powiązania między zmiennymi (zjawiskami).**

***Zmienna objaśniana (zależna):* zjawisko ekonomiczne wyjaśniane przez model.**

***Zmienna objaśniająca (niezależna):* zjawisko ekonomiczne oddziałujące (wpływające) na zmienną objaśnianą.**

MODEL LINIOWEJ REGRESJI DWUWYMIAROWEJ

***Funkcja regresji I rodzaju*: funkcja przyporządkowująca wartości zmiennej objaśniającej oczekiwaną wartość zmiennej objaśnianej.**

$$\hat{Y} = E(Y/X = x) = \alpha_0 + \alpha_1 X \quad \text{lub} \quad \hat{X} = E(X/Y = y) = \beta_0 + \beta_1 Y$$

***Funkcja regresji II rodzaju*: funkcja dla n-elementowej próby otrzymana metodą najmniejszych kwadratów (MNK):**

$$\hat{y}_i = a_0 + a_1 x_i \quad \text{lub} \quad \hat{x}_i = b_0 + b_1 y_i \quad i = 1,2,3,...,n$$

gdzie: x_i, y_i - realizacje wartości zmiennych X i Y w próbie (empiryczne),

$\hat{x}_i, \hat{y}_i$ - wartości zmiennej X i Y wyznaczone na podstawie funkcji II rodzaju (teoretyczne),

a_0, b_0 - wyrazy wolne funkcji regresji II

a_1, b_1 - współczynniki funkcji regresji II rodzaju.

ANALIZA REGRESJI

Konkretne wartości a_0, b_0, a_1, b_1 (np. 1, -2, 3 ,9) są ocenami parametrów α_0, β_0, α_1, β_1.

Do oceny dopasowania wyznaczonej funkcji regresji II rodzaju do badanej rzeczywistości wykorzystuje się tzw. reszty - błąd zwany składnikiem resztowym *e (różnica pomiędzy wartościami empirycznymi a teoretycznymi funkcji regresji)*.

Dla regresji Y względem X (funkcji $\hat{y}_i = a_0 + a_1 x_1$)

$$e_i = y_i - \hat{y}_i$$

gdzie: e_i - składnik resztowy,
y_i - wartości empiryczne,
$\hat{y}_i$ - wartości teoretyczne.

Dla regresji X względem Y (funkcji $\hat{x}_i = b_0 + b_1 y_1$)

$$z_i = x_i - \hat{x}_i$$

gdzie: e_i - składnik resztowy,
x_i - wartości empiryczne,
$\hat{x}_i$ - wartości teoretyczne.

Przyczyny występowania składnika resztowego to:

- **zdarzenia losowe,**
- **błędna analityczna postać funkcji regresji,**
- **pominięcie ważnych zmiennych objaśniających,**
- **inne (np. błędy w pomiarze zmiennych).**

ANALIZA REGRESJI

ESTYMACJA PARAMETRÓW FUNKCJI REGRESJI I RODZAJU

Estymację parametrów funkcji regresji I rodzaju $\hat{Y} = \alpha_0 + \alpha_1 X$ **przeprowadza się <u>klasyczną metodą najmniejszych kwadratów (KMNK),</u> która polega na takim ich oszacowaniu aby dla n obserwacji spełniony był warunek :**

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a_0 - a_1 x_1)^2 \Rightarrow \min$$

Warunkiem koniecznym aby takie minimum istniało jest zerowanie się pochodnych cząstkowych funkcji w względem a_0 i a_1.

$$\frac{\partial(W)}{\partial(a_0)} = -2(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_0 - \sum_{i=1}^{n} a_1 x_i) = 0 \qquad \frac{\partial(W)}{\partial(a_1)} = -2(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} a_0 x_i - \sum_{i=1}^{n} a_1 x_i^2) = 0$$

Po odpowiednich przekształceniach otrzymujemy układ równań o postaci:

$$\sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2$$

z czego:

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - a_1 \bar{x}$$

ANALIZA REGRESJI

ESTYMACJA PARAMETRÓW FUNKCJI REGRESJI I RODZAJU

Jeżeli zamiast wartości zmiennych x_i i y_i użyjemy ich odchyleń $(x_i - \bar{x})$ i $(y_i - \bar{y})$

to elementy układu równań postaci: $\sum_{i=1}^{n}(x_i - \bar{x})$ i $\sum_{i=1}^{n}(y_i - \bar{y})$ **zerują się i wówczas:**

$$a_0 = 0, \text{ oraz } a_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Wartość oceny parametru a_1 określa, o ile jednostek przeciętnie wzrośnie (lub spadnie) wartość zmiennej zależnej, gdy wartość zmiennej niezależnej wzrośnie o jedną jednostkę.

ANALIZA REGRESJI

MIERNIKI DOPASOWANIA FUNKCJI REGRESJI DO DANYCH EMPIRYCZNYCH

Funkcja regresji jest poprawnie oszacowana, jeżeli wartości reszt są niewielkie i mają charakter losowy.

WARIANCJA SKŁADNIKA RESZTOWEGO

Dla regresji Y względem X stosuje się wzór:

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

gdzie: k - liczba zmiennych objaśniających

W celach interpretacyjnych posługujemy się odchyleniem standardowym składnika resztowego S_e, które mówi, jaki średnio błąd popełnia się dopasowując funkcję regresji do danych empirycznych.

ANALIZA REGRESJI

MIERNIKI DOPASOWANIA FUNKCJI REGRESJI DO DANYCH EMPIRYCZNYCH

WSPÓŁCZYNNIK DETERMINACJI r^2 (R^2) i WSPÓŁCZYNNIK ZBIEŻNOŚCI (indeterminacji) $\varphi 2$

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad \varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad 1 = r^2 + \varphi^2$$

… czyli 1 - φ^2 = r^2. Obydwie wielkości r^2 , jak i φ^2 można wyrazić w procentach. Interpretując r^2, określa się współczynnik determinacji jako procent wyjaśnienia zmienności zmiennej zależnej (objaśnianej) przez zmienność zmiennej niezależnej (objaśniającej).

ANALIZA REGRESJI

MIERNIKI DOPASOWANIA FUNKCJI REGRESJI DO DANYCH EMPIRYCZNYCH

ŚREDNIE BŁĘDY SZACUNKU PARAMETRÓW FUNKCJI REGRESJI LINIOWEJ (dwuwymiarowej)

Średnie błędy szacunku parametrów funkcji regresji informują nas o tym, o ile średnio mylimy się (*in plus* lub *in minus*), gdy szacujemy parametry α_0 i α_1 w populacji generalnej na podstawie informacji zaczerpniętych z próby losowej.

***Średnie błędy szacunku parametrów powinny być bliskie zera, by świadczyć o dokładnym oszacowaniu tych parametrów*.**

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

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

Przyczyny dużych średnich błędów szacunku parametrów:

1) **Mała liczebność próby (im n większe tym S^2_e mniejsze),**
2) **Niewłaściwa metoda estymacji parametrów funkcji regresji,**
3) **Niewłaściwy wybór zmiennej objaśniającej do funkcji regresji.**