Metody prognozowania:

Regresja liniowa i nieliniowa

Dr inż. Sebastian Skoczypiec

Definicja: *Próbą* (ang. *sample*) nazywamy skończony podzbiór populacji, który poddajemy badaniu ze względu na interesującą nas cechę.

Przykłady:

- populacja: przedsiębiorstwa zarejestrowane w Polsce, próba: przedsiębiorstwa województwa kujawsko-pomorskiego,
- populacja: przedsiębiorstwa województwa kujawsko-pomorskiego, próba: wybrane 20 przedsiębiorstw.

Próba powinna reprezentować populację w tym sensie, że częstość występowania każdej z badanych cech w próbie nie powinna się różnić od częstości występowania tej cechy w całej populacji.

Próby obciążone — uzyskiwane np. przez wywiad telefoniczny, czy ankietowe badania internetowe.

- *Próba losowa prosta* — każdy element populacji ma jednakową szansę znalezienia się w próbie (umiemy określić liczbowo jaka to szansa), wybierana najczęściej z użyciem liczb losowych.
- *Próba systematyczna* — ze spisu elementów populacji wybieramy co n-ty.
- *Próba kwotowa* — częsta w badaniach rynku, ankieter wybiera dowolne osoby posiadające określone cechy np. 5 mężczyzn powyżej 60. roku życia (dowolnych), 2 gospodynie domowe (dowolne), 3 studentki (dowolne) itp.

- *Statystyka opisowa* (ang. *descriptive statistics*) zajmuje się prezentacją danych w postaci tabel, diagramów i charakterystyk liczbowych.
- *Statystyka matematyczna* (ang. *mathematical* lub *inductive statistics*) zajmuje się wnioskowaniem o własnościach populacji na podstawie własności próbki przy dopuszczeniu pewnego poziomu błędu, w oparciu o twierdzenia rachunku prawdopodobieństwa.

Miary tendencji centralnej

Oznaczenia:

- N — liczebność próbki,
- $x_1, x_2, \ldots, x_N$ — obserwacje,
- $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ — obserwacje ustawione niemalejąco.

Miary tendencji centralnej:

- średnia (ang. *mean*),
- mediana (ang. *median*),
- moda, inaczej dominanta (ang. *mode*).

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

Średnią podajemy z dokładnością o 1 większą niż dane.

- Suma odchyleń wszystkich wartości zmiennej od średniej jest równa 0.
- Suma kwadratów odchyleń wartości zmiennej od pewnej liczby a jest najmniejsza dla a będącego średnią.

- Zalety:
  - Może być wykorzystywana w dalszych obliczeniach statystycznych.
  - Jest najmniej podatna na błąd jako przybliżenie średniej dla całej populacji.
- Wady:
  - Wrażliwa na nienormalnie duże lub nienormalnie małe wartości skrajne.
  - W przypadku rozkładów dwu- i wielomodalnych bywa myląca.

- Przykład 1.

$$12, 36, 18, 25, 24, 11, 39, 11, 29, 35.$$

Średnia podanych liczb to $24,0$.

- W dowcipie rysunkowym robotnik mówi do dziennikarki: *Średnio rocznie w naszej firmie zarabia się* $100\,000$ *zł. Prezes zarabia milion, a nasza dziesiątka po* $10\,000$.

$$\frac{1\,000\,000 + 10 \cdot 10\,000}{11} = \frac{1\,100\,000}{11} = 100\,000.$$

Mediana

- Wartość środkowa. Jeśli N jest nieparzyste, to medianą jest $x_{((N+1)/2)}$, a jeśli parzyste, to $\frac{x_{(N/2)}+x_{((N/2)+1)}}{2}$.
- Suma odchyleń bezwzględnych od mediany jest mniejsza niż suma takich odchyleń od jakiejkolwiek innej liczby.
- W celu obliczenia mediany dane z przykładu 1. porządkujemy:

$$11, 11, 12, 18, 24, 25, 29, 35, 36, 39.$$

- Zalety:
  - Łatwa do zrozumienia.
  - Nie ulega deformacji ze względu na nienormalnie duże lub nienormalnie małe wartości skrajne.
- Wady:
  - Nie może być wykorzystywana w dalszych obliczeniach statystycznych.
  - Dla małych zbiorów danych, o pewnej szczególnej postaci, nie jest dobrą charakterystyką tendencji centralnej (np. medianą dla 5, 5, 5, 9, 10 jest 5).

Miary rozproszenia

- Rozstęp (ang. *range*) $R = x_{max} - x_{min}$.
- Kwantyle(ang. *quantiles*):
  - kwartyle (ang. *quartiles*),
  - decyle (ang. *deciles*) — Sir Francis Galton (1882),
  - percentyle (ang. *percentiles*) — Sir Francis Galton (1885).
- Odchylenie standardowe (ang. *standard deviation*) — Karl Pearson (1893).

Odchylenie standardowe

$$\hat{s} = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N}} = \sqrt{\frac{x_1^2 + \ldots + x_N^2}{N} - \bar{x}^2}.$$

W przypadku, gdy zgromadzone dane traktujemy jako dane całej populacji, odchylenie standardowe obliczamy, dzieląc powyższe sumy przez N. Jeśli natomiast analizujemy próbkę i otrzymane odchylenie standardowe ma być przybliżeniem odchylenia standardowego w całej populacji, należy dzielić przez $N - 1$ (tak liczą programy statystyczne). Zapobiega to obciążeniu tego przybliżenia (estymatora).

Własności odchylenia standardowego

- Jeżeli do wszystkich wartości zmiennej dodamy pewną wartość stałą, to odchylenie standardowe nie zmienia się.
- Jeżeli wszystkie wartości zmiennej pomnożymy przez pewną liczbę, to odchylenie standardowe również zostanie pomnożone przez tą liczbę.
- Odchylenie standardowe może nie być dobrą miarą rozproszenia, gdy zmienna przyjmuje kilka wartości bardzo oddalonych od reszty lub gdy rozkład jest mocno skośny.

Zmienna losowa

- **Zmienna losowa X** – zmienna, która w wyniku pewnego doświadczenia przyjmuje z pewnym prawdopodobieństwem wartość z określonego zbioru
- **Zmienną losową X** – nazywamy **dyskretną (skokową)**, jeżeli zbiór wartości zmiennej X jest zbiorem skończonym lub przeliczalnym (ciąg liczbowy).
- **Zmienną losową X** – nazywamy **ciągłą**, jeżeli zbiór wartości zmiennej X można przedstawić jako przedział liczbowy.

Zmienna losowa skokowa

- Rozkładem zmiennej losowej skokowej (funkcją rozkładu prawdopodobieństwa) nazywamy funkcję prawdopodobieństwa, która każdej realizacji zmiennej X przyporządkowuje określone prawdopodobieństwo:

$$P(X = x_i) = p_i$$

- dla pi>=0
- gdzie: P(X=xi) – prawdopodobieństwo, że zmienna X przyjmie wartość xi,
- $\sum_{i=1}^{\infty} p_i = 1$
- *Dystrybuantą zmiennej losowej X* nazywamy funkcję F(x) dla wszystkich liczb rzeczywistych o postaci

$$F(x) = P(X \le x) = \sum_{x_i \le x} p_i$$

$$F(x) = \begin{cases} 0 & dla\ x < x_1 \\ p_1 & dla\ x_1 \le x < x_2 \\ p_1 + p_2 & dla\ x_2 \le x < x_3 \\ \vdots & \\ 1 & dla\ x \ge x_{i-1} \end{cases}$$

Zmienna losowa ciągła

- Funkcją gęstości prawdopodobieństwa zmiennej losowej ciągłej nazywamy funkcję *f(x)*, określoną na zbiorze liczb rzeczywistych, spełniającą warunki:

$$f(x) \geq 0 \qquad \int_a^b f(x)dx = P(a \leq X \leq b) = P(a < X < b)$$

- dla dla dowolnych *a* < *b*

$$\sum_{i=1}^{\infty} p_i = 1 \qquad \int_{-\infty}^{\infty} f(x)dx = 1 \qquad \int_a^b f(x)dx = 1 \qquad P(X = a) = 0$$

- *Dystrybuantą zmiennej losowej X* ciągłej nazwyamy funkcję:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$$

$$P(X < x_1) = F(x_1)$$

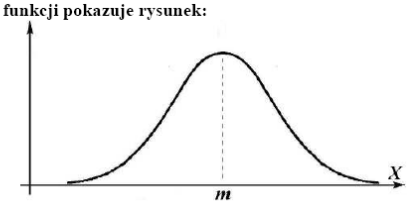
$$P(x_2 < X < x_3) = F(x_3) - F(x_2)$$

$$P(X > x_4) = 1 - F(x_4)$$

Zmienna losowa X (ciągła) ma rozkład normalny z parametrami *m* i σ jeśli jej funkcja gęstości jest określona wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad \text{dla} \quad -\infty < x < +\infty$$

Wykres tej funkcji pokazuje rysunek:



Rozkład zmiennej losowej X ma następujące parametry:

wartość oczekiwana $E(X) = m$

odchylenie standardowe $\sqrt{V(X)} = \sigma$

Rozkład chi-kwadrat (rozkład χ^2)

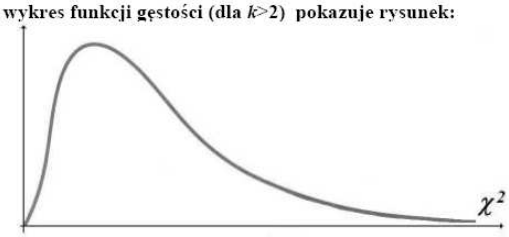
Danych jest k ciągłych zmiennych losowych o rozkładzie normalnym z wartością oczekiwaną 0 i odchyleniem standardowym 1, tj. każda zmienna $X_i : N(0;1)$ (i=1,2, ... ,k).

Zdefiniujemy nową zmienną losową o nazwie chi-kwadrat (χ^2) :

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_k^2$$

Rozkład tak zdefiniowanej zmiennej χ^2 nazywamy rozkładem zmiennej losowej chi-kwadrat o k stopniach swobody.

Typowy wykres funkcji gęstości (dla k>2) pokazuje rysunek:



Rozkład t - Studenta

Dana są dwie zmienne losowe:

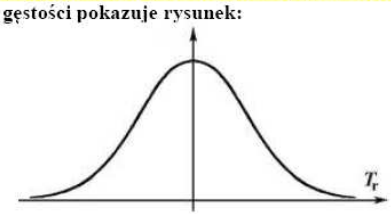
1. zmienna losowa $U:N(0;1)$ oraz
2. zmienna losowa χ^2 o k stopniach swobody

Zdefiniujemy nową zmienną losową postaci :

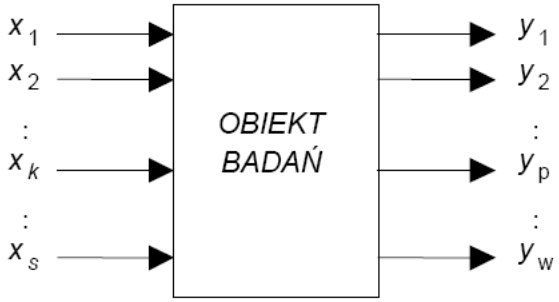
$$T_k = \frac{U}{\sqrt{\chi^2}}\sqrt{k}$$

Rozkład tak zdefiniowanej zmiennej T_k nazywamy rozkładem zmiennej losowej t-Studenta o k stopniach swobody.

Wykres funkcji gęstości pokazuje rysunek:



Wprowadzenie



Wprowadzenie

Korelacja:
rodzaj zależności pomiędzy zmiennymi losowymi, z których każda wyznaczona jest przez pewną cechę, ze względu na którą bada się dano populację.

Regresja:
sprowadzenie zagadnienia współzależności zmiennych losowych do zależności funkcyjnej.

Na podstawie wyników badań doświadczalnych wyznacza się zależność pomiędzy zmiennymi losowymi, najczęściej w formie tzw. równania regresji, które przedstawia charakter związków pomiędzy ***czynnikami wejściowymi i wynikowymi.***

Z matematycznego punktu widzenia, regresją nazywamy dowolną metodę statystyczną pozwalającą estymować warunkową wartość oczekiwaną zmiennej losowej, ***zwanej zmienną objaśnianą****, dla zadanych wartości innej zmiennej lub wektora zmiennych losowych (tzw.* ***zmiennych objaśniających****).*

Regresja

Definicja: *Regresją* nazywamy wyrażenie zależności wartości jednej zmiennej od drugiej w postaci pewnej prostej funkcji z dopuszczeniem ewentualnych odstępstw. Pierwsze wyniki: Sir Francis Galton *Regression towards Mediocrity in Hereditary Stature* (*Regresja w badaniach nad dziedziczeniem niskiego wzrostu*), 1885 r. Teoria została później rozwinięta przez Karla Pearsona.

Wprowadzenie

Opis zależności cechy Y od cechy X (regresji cechy Y względem cechy X) – terminologia

cecha X : objaśniająca, opisująca, niezależna

cecha Y : objaśniana, opisywana, zależna

W zapisie formalnym model przybiera postać:

$$Y = f(X,\beta) + \varepsilon \quad \text{lub} \quad Y = f(X+ \varepsilon_X,\beta) + \varepsilon$$

gdzie: *X – wektor zmiennych objaśniających*
Y – zmienna objaśniana
β - wektor współczynników regresji
$f(X,\beta)$ – funkcja regresji
ε, ε_X - błąd losowy

Wprowadzenie

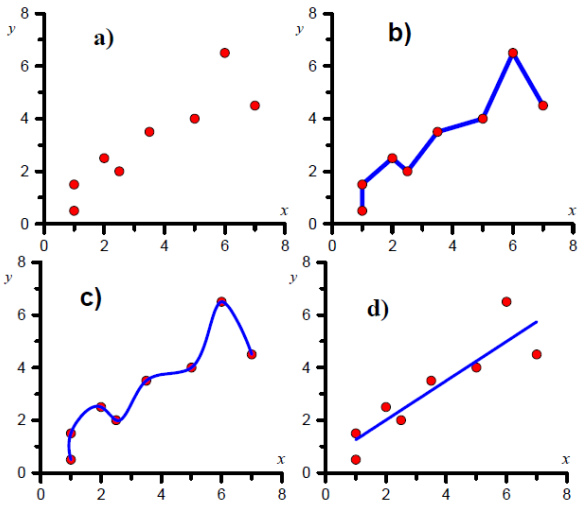
Celem konstrukcji modelu jest przybliżenie nieznanej funkcji f przez jej estymator. Sprowadza się to do takiego wyznaczenia wektora współczynników β, aby zminimalizować w zbiorze uczącym **funkcję straty.**

$$L(f\,,\ \bar{f}) = f(\Delta(a,b))$$

Zwykle jako miarę błędów stosuje się sumę kwadratów różnic (błędów regresji):

$$\Delta(a,b) = \Delta(a-b)^2$$

wówczas obliczenia są najprostsze - dopasowanie modelu sprowadza się do zastosowania prostej matematycznie metody najmniejszych kwadratów (MNK).



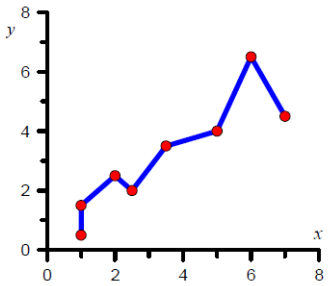
a) przykładowe wyniki pomiarów

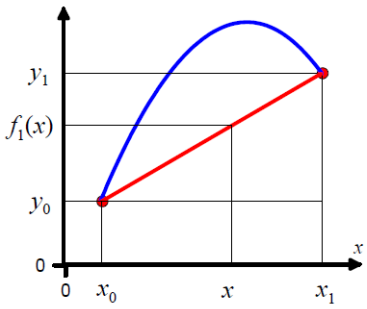
b) interpolacja liniowa

c) interpolacja krzywoliniowa

d) aproksymacja liniowa

INTERPOLACJA LINIOWA

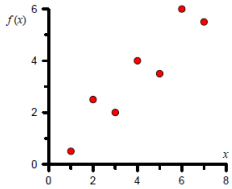


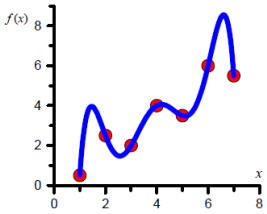


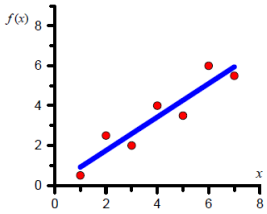
$$\frac{f_1(x)-y_0}{x-x_0}=\frac{y_1-y_0}{x_1-x_0}$$

$$f_1(x)=y_0+\frac{y_1-y_0}{x_1-x_0}(x-x_0) \qquad (1.1)$$

Podsumowanie: interpolacja polegała na sformułowaniu równania, które będzie przechodzić przez wszystkie znane punkty. Jeżeli dane (punkty) wyznaczone są z błędami (a takimi zazwyczaj są dane pomiarowe) takie postępowanie prowadzi do sfałszowania obrazu.







Zbiór siedmiu punktów oraz interpolacja wielomianem (6 stopnia) oraz aproksymacja prostą

Badamy jednostki statystyczne pod kątem dwóch różnych cech - cechy X oraz cechy Y.
Pytanie jakie sobie stawiamy to:

czy istnieje zależność pomiędzy cechą X i cechą Y ?

Jeżeli taka zależność istnieje, to poszukujemy odpowiedzi na kolejne pytania:

- jaki jest charakter tej zależności oraz
- jaka jest jej siła ?

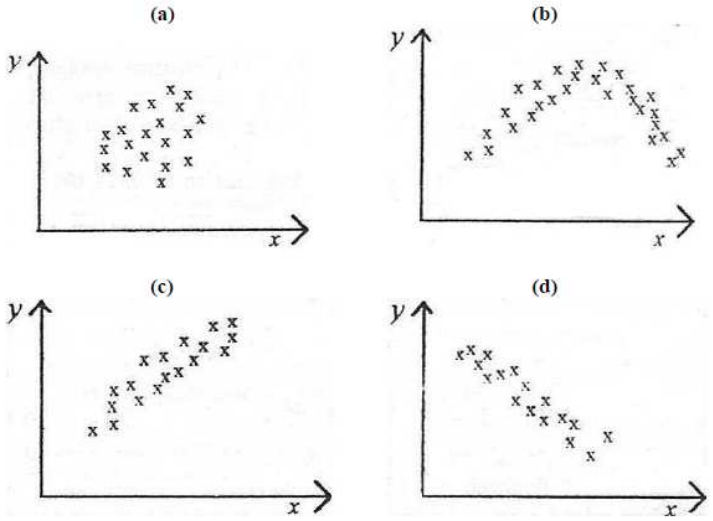
Zależność korelacyjna pomiędzy cechami X i Y charakteryzuje się tym, że wartościom jednej cechy są przyporządkowane ściśle określone wartości średnie drugiej cechy.

Korelacyjny wykres rozrzutu KORELOGRAM

Jeżeli obie cechy X i Y są mierzalne, to analizę zależności rozpoczynamy od sporządzenia korelogrmamu.
Korelogram jest to wykres punktowy par { (x_i , y_i) }.
(Excel nazywa taki wykres: „wykresem XY”).
W kartezjańskim układzie współrzędnych $x0y$ pary te odpowiadają punktom o współrzędnych

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

PRZYKŁADY korelogramów (*każdy punkt oznaczono* x)



Jeżeli otrzymamy <u>bezładny zbiór punktów</u>, który nie przypomina kształtem wykresu znanego związku funkcyjnego, to powiemy że pomiędzy cechami X i Y <u>nie ma zależności</u>. Ilustruje to rysunek (a).

Na rysunku (b) widać, że smuga punktów układa się w kształt paraboli. Powiemy zatem, że istnieje zależność pomiędzy cechami X i Y i jest to związek nieliniowy; <u>zależność nieliniowa</u>.

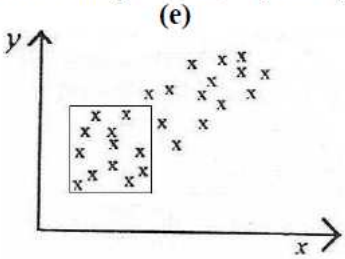
Na rysunkach (c) i (d) smuga punktów układa się wzdłuż <u>linii prostej</u>. Powiemy zatem, że istnieje zależność pomiędzy cechami X i Y i jest to związek liniowy; <u>zależność liniowa</u>.

Rysunki (e) i (f) ilustrują przypadki <u>błędów we wnioskowaniu</u> o zależności cech X i Y na podstawie korelogramu.

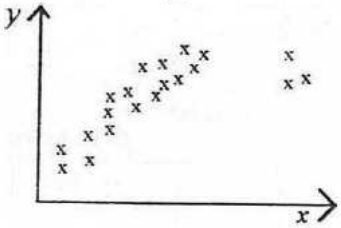
Rysunek (e) – <u>za mało danych</u>. Zebrano dane (punkty obwiedzione kwadratem) i z korelogramu wynika brak zależności. W rzeczywistości jest zależność liniowa.

Rysunek (f) – <u>nietypowe dane</u>. Trzy ostatnie punkty (odseparowane) to dane nietypowe. Sugerują zależność nieliniową (parabola). Po odrzuceniu tych nietypowych informacji widać, że jest wyraźna zależność liniowa.

(e)



(f)



Pomiar KIERUNKU i SIŁY zależności <u>liniowej</u>
Szeregi <u>szczegółowe</u>

WSPÓŁCZYNNIK KORELACJI (Pearsona)

Współczynnik korelacji (Pearsona) r_{xy} obliczamy dla cech <u>ilościowych</u> wg następującego wzoru:

$$r_{xy} = \frac{C(X,Y)}{s_x s_y}$$

gdzie:

$C(X,Y)$ – <u>kowariancja</u> pomiędzy cechami X i Y

s_x (s_y) – odchylenie standardowe cechy X (cechy Y)

Kowariancja jest kluczowym parametrem rozkładu dwóch cech w badaniu zależności cech ilościowych X i Y. Wylicza się ją wg następującego wzoru (dla szeregu(ów) szczegółowego):

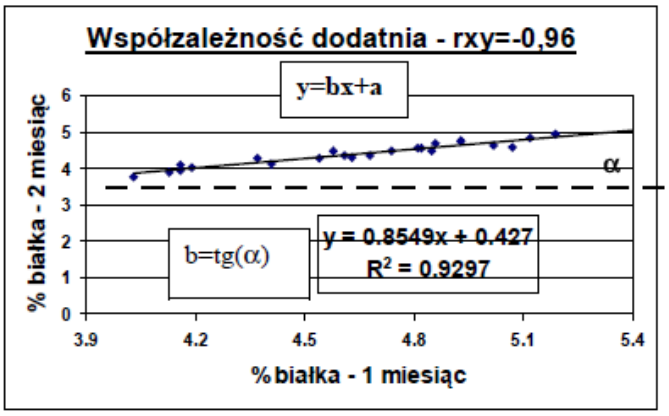
$$C(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

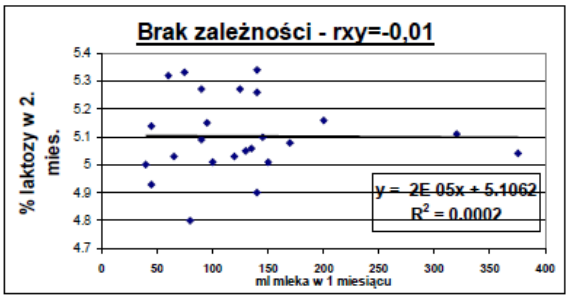
Współczynnik korelacji (Pearsona) r_{xy} **spełnia zawsze warunek:**

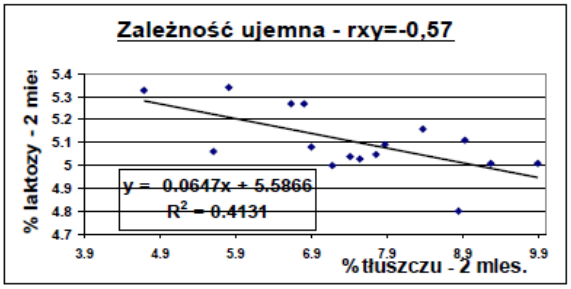
$$\boxed{-1 \le r_{xy} \le 1}$$

Współczynnik korelacji (Pearsona) jest miarą symetryczną, tzn.

$$r_{xy} = r_{yx}$$







r_{xy} = jest liczbą niemianowaną, przyjmującą wartości od -1 do +1.
r_{xy} = -1 lub +1 całkowita zależność cech X i Y
r_{xy} = 0 brak jakiejkolwiek zależności

W statystyce stopień zależności między cechami można wyrazić wg następującej skali:

Skala Stanisza

$r_{xy} = 0$	zmienne nie są skorelowane
$0 < r_{xy} \leq 0,1$	korelacja nikła
$0,1 < r_{xy} \leq 0,3$	korelacja słaba
$0,3 \leq r_{xy} \leq 0,5$	korelacja przeciętna
$0,5 \leq r_{xy} \leq 0,7$	korelacja wysoka
$0,7 \leq r_{xy} \leq 0,9$	korelacja bardzo wysoka
$0,9 \leq r_{xy} \leq 1$	korelacja prawie pewna

Skala Guillforda

Przedział	Zależność	Współczynnik
0,00±0,20	Słaba	Prawie nic nieznaczący
±0,20±040	Niska	Wyraźna, ale słaba
±0,40±0,70	Umiarkowana	Rzeczywisty
±0,70±0,90	Wysoka	Znaczny
±0,90±1,00	Bardzo wysoka	Pewny

Znak współczynnika r_{xy} mówi nam o kierunku zależności. I tak:

- znak *plus* – zależność liniowa dodatnia, tzn. wraz ze wzrostem wartości jednej cechy rosną średnie wartości drugiej z cech,
- znak *minus* – zależność liniowa ujemna, tzn. wraz ze wzrostem wartości jednej cechy maleją średnie wartości drugiej z cech.

Wartość bezwzględna współczynnika korelacji, czyli $|r_{xy}|$, mówi nam o sile zależności. Jeżeli wartość bezwzględna $|r_{xy}|$:

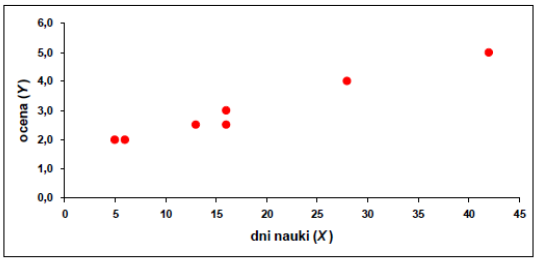
- jest mniejsza od 0,2, to praktycznie brak związku liniowego pomiędzy badanymi cechami,
- 0,2 – 0,4 - zależność liniowa wyraźna, lecz niska,
- 0,4 – 0,7 - zależność liniowa umiarkowana,
- 0,7 – 0,9 - zależność liniowa znacząca,
- powyżej 0,9 - zależność liniowa bardzo silna.

PRZYKŁAD 1

W grupie 7 studentów badano zależność pomiędzy oceną z egzaminu ze statystyki (Y), a liczbą dni poświęconych na naukę (X).

nr studenta	*ocena z egzaminu* (*Y*)	*liczba dni nauki* (*X*)
i	y_i	x_i
1	2,0	5
2	2,5	13
3	2,5	16
4	4,0	28
5	5,0	42
6	3,0	16
7	2,0	6

Sporządzamy korelogram.



Widać tutaj wyraźną zależność liniową (dodatnią).
Obliczamy współczynnik korelacji (Pearsona).
UWAGA ! Liczebność populacji jest mała (n=7). Użyjemy tak małego przykładu tylko dlatego, aby sprawnie zilustrować procedurę liczenia.

Obliczanie średnich, wariancji oraz kowariancji.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	(4)*(4)	(5)*(5)	(4)*(5)
1	2,0	5					
2	2,5	13					
3	2,5	16					
4	4,0	28	1,0	10	1,00	100	10,0
5	5,0	42	2,0	24	4,00	576	48,0
6	3,0	16	0,0	-2	0,00	4	0,0
7	2,0	6	-1,0	-12	1,00	144	12,0
razem			X	X			

$$n = 7 \qquad \bar{x} = \frac{126}{7} = 18 \qquad \bar{y} = \frac{21}{7} = 3$$

$$s_x^2 = \frac{1022}{7} = 146 \qquad s_y^2 = \frac{7,5}{7} = 1,07$$

$$s_x = \sqrt{146} = 12,08 \qquad s_y = \sqrt{1,07} = 1,03$$

$$C(X,Y) = \frac{86,5}{7} = 12,36$$

Współczynnik korelacji (Pearsona) wynosi dla danych z przykładu 1:

$$r_{xy} = \frac{C(X,Y)}{s_x s_y} = \frac{12,36}{12,08 \times 1,03} = +0,993$$

INTERPRETACJA

W badanej grupie studentów wystąpiła bardzo silna dodatnia (znak *plus*) zależność liniowa pomiędzy czasem nauki (cecha X), a uzyskaną oceną z egzaminu (cecha Y).

Oznacza to, że wraz ze wzrostem czasu poświęconego na naukę rosła w tej grupie uzyskiwana ocena.

WSPÓŁCZYNNIK KORELACJI RANG
(Spearmana)

Współczynnik korelacji rang (Spearmana) r_S używamy w przypadku gdy:

1. **choć jedna z badanych cech jest cechą jakościową (niemierzalną), ale istnieje możliwość uporządkowania (ponumerowania) wariantów każdej z cech;**
2. **cechy mają charakter ilościowy (mierzalny), ale liczebność zbiorowości jest mała (n<30).**

Numery jakie nadajemy wariantom cech noszą nazwę rang.

Współczynnik korelacji rang (Spearmana) r_S wyznaczamy wg następującego wzoru:

$$r_S = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

d_i – różnica pomiędzy rangami dla cechy X i cechy Y

Współczynnik korelacji rang (Spearmana) r_S spełnia zawsze warunek:

$$\boxed{-1 \le r_S \le 1}$$

INTERPRETACJA
Analogiczna jak dla współczynnika korelacji (Pearsona).

TABLICA KORELACYJNA

Schemat tablicy korelacyjnej

Warianty cechy X (x_i)	Warianty cechy Y (y_j)				(razem) $n_{i\bullet}$
	y_1	y_1	$\cdots$	y_s	
x_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\bullet}$
(razem) $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet s}$	n

Oznaczenia:

n_{ij} - **liczba jednostek, która charakteryzuje się wartością x_i cechy X oraz wartością y_j cechy Y**

$n_{i\bullet}$ - **liczba jednostek, która charakteryzuje się wartością x_i cechy X**

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$$

$n_{\bullet j}$ - **liczba jednostek, która charakteryzuje się wartością y_j cechy Y**

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$$

n - **liczebność populacji**

$$n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j}$$

PRZYKŁAD 3

Podobnie jak w przykładzie 1 zbadamy zależność pomiędzy czasem nauki (X), a uzyskaną oceną (Y).

W tablicy korelacyjnej zestawiono informację o 400 studentach (n=400).

Ocena (Y)	Czas nauki (X) *w dniach*				$n_{i\bullet}$
	0 - 7	7 - 14	14 - 21	21 - 28	
2	80				
3	10	80			
3,5		60	10		
4		20	30		
4,5			50	10	
5				50	
$n_{\bullet j}$					

Obliczamy <u>osobno</u> dla każdej z cech: średnie, wariancje i odchylenia standardowe.

Ocena (Y)	Czas nauki (X)				(a)	$y_i n_{i\bullet}$	(b)	(c)	(d)
	0 - 7	7 - 14	14 - 21	21 - 28	$n_{i\bullet}$		$y_i - \bar{y}$	(b)*(b)	(c)*(a)
2	80				80				
3	10	80			90				
3,5		60	10		70				
4		20	30		50	200	0,5	0,25	12,5
4,5			50	10	60	270	1	1	60
5				50	50	250	1,5	2,25	112,5
$n_{\bullet j}$	90	160	90	60	400		x	x	
$\dot{x}_j$			17,5	24,5	x	x	x	x	x
$\dot{x}_j n_{\bullet j}$			1575	1470		x	x	x	x
$\dot{x}_j - \bar{x}$			4,9	11,9	x	x	x	x	x
$(\dot{x}_j - \bar{x})^2$			24,01	141,61	x	x	x	x	x
$(\dot{x}_j - \bar{x})^2 n_{\bullet j}$			2160,9	8496,6		x	x	x	x

$$n = 400$$

$$\bar{x} = \frac{5040}{400} = 12{,}6 \qquad \bar{y} = \frac{1395}{400} = 3{,}5$$

$$s_x^2 = \frac{18816}{400} = 47{,}04 \qquad s_y^2 = \frac{387{,}5}{400} = 0{,}97$$

$$s_x = \sqrt{47{,}04} = 6{,}86 \qquad s_y = \sqrt{0{,}97} = 0{,}98$$

Przechodzimy do obliczania kowariancji *C(X,Y)*.

Na początek policzymy wszystkie iloczyny $(\dot{x}_j - \bar{x})(y_i - \bar{y})$:

$y_i - \bar{y}$	$\dot{x}_j - \bar{x}$			
	-9,1	-2,1	4,9	11,9
-1,5				
-0,5				
0	0	0	0	0
0,5	-4,55	-1,05	2,45	5,95
1	-9,1	-2,1	4,9	11,9
1,5	-13,65	-3,15	7,35	17,85

Wykorzystamy tabelę początkową:

Ocena (Y)	Czas nauki (X) *w dniach*			
	0 - 7	7 - 14	14 - 21	21 - 28
2	80	0	0	0
3	10	80	0	0
3,5	0	60	10	0
4	0	20	30	0
4,5	0	0	50	10
5	0	0	0	50

i policzymy wszystkie iloczyny $(\dot{x}_j - \bar{x})(y_i - \bar{y})n_{ij}$

$y_i - \bar{y}$	$\dot{x}_j - \bar{x}$				*razem*
	-9,1	-2,1	4,9	11,9	
-1,5					
-0,5					
0	0	0	0	0	
0,5	0	-21	73,5	0	
1	0	0	245	119	
1,5	0	0	0	892,5	
razem					

Zatem kowariancja wynosi:

$$C(X,Y) = \frac{2530{,}5}{400} = 6{,}33$$

Współczynnik korelacji (Pearsona) wynosi dla danych z przykładu 3:

$$r_{xy} = \frac{C(X,Y)}{s_x s_y} = \frac{6{,}33}{6{,}86 \times 0{,}98} = +0{,}942$$

INTERPRETACJA
W badanej grupie 400 studentów wystąpiła bardzo silna dodatnia (znak *plus*) zależność liniowa pomiędzy czasem nauki (cecha X), a uzyskaną oceną z egzaminu (cecha Y).

Inne miary zależności
wyliczalne na podstawie tablicy korelacyjnej

Obok współczynnika korelacji Persona stosowane są inne miary zależności pomiędzy cechą Y i cechą X. Są to:

- **Stosunek korelacji (e_{yx})**
- **Miary oparte na chi-kwadrat (χ^2)**

Stosunek korelacji

- Miara ta jest oparta na spostrzeżeniu, że przy braku zależności średnie poziomy cechy Y wewnątrz grup (klas) pokrywają się ze średnią ogólną cechy Y.
- Miara ta spełnia warunki

$$0 < e_{yx} < 1$$

$$|r_{yx}| \leq e_{yx}$$

- Warunkiem policzenia stosunku korelacji jest mierzalność cechy Y.
- Jest to miara zalecana w przypadku badania zależności dla związków nieliniowych.

Stosunek korelacyjny

Współczynnik korelacji r nie jest czuły na zależności krzywoliniowe.

Gdy zależność między dwoma zmiennymi jest nieliniowa, wówczas miara koncentracji wyników pomiarów względem krzywej regresji może być tzw. **stosunek korelacyjny:**

$$\eta_{x|y} = \sqrt{1 - \frac{\left[\sum_{j=1}^{k}(m_j - 1)S^2_{x|y_j}\right]^2}{(n-1)S^2_y}}$$

gdzie: k – liczba przedziałów, $S^2_{x|y_j}$ - wariancja dla j przedziału, m_j – ilość punktów w j przedziale

Stosunek korelacyjny

Stosunek korelacyjny: określa stosunek pomiędzy dwoma zmiennymi, których zależność przyczynowo skutkowa jest określona (x zależy od y).

Jeżeli zależność ta nie jest znana to należy określić $\eta_{x|y}$.

$\eta_{x|y}$ = 0: brak koleralcji między badanymi zmiennymi (tzn. brak zależności zmiennej y od x)

$\eta_{x|y}$ = 1: zależność pomiędzy x i y jest funkcyjna

$\eta_{x|y}$ = $r_{x|y}$: zależność liniowa

Charakter relacji

Współczynnik korelacji liniowej	Stosunek korelacji	Zależność między zmiennymi x i y
$r_{x\|y}=\pm1$	-	funkcyjna liniowa
$r_{x\|y}=0$	$\eta_{x\|y}=1$	funkcja krzywoliniowa
$r_{x\|y}=0$	$\eta_{x\|y}=0$	brak korelacji
$r_{x\|y}=0$	$\eta_{x\|y}<1$	korelacja krzywoliniowa
$r_{x\|y}=\eta_{x\|y}$	-	dokładna korelacja liniowa
$\|r_{x\|y}\|>0$	$\eta_{x\|y}<1$	korelacja liniowa

Rozważymy tylko przypadek zależności liniowej dla regresji prostej. Narzędziem będzie zatem funkcja regresji postaci:

$$\hat{y}_i = ax_i + b$$

$\hat{y}_i$ **- teoretyczna wartość zmiennej zależnej (Y)**

x_i **- empiryczna wartość zmiennej niezależnej (X)**

a **– współczynnik regresji (współczynnik kierunkowy)**

INTERPRETACJA: jeżeli wartość zmiennej niezależnej X wzrośnie o jednostkę, to wartość zmiennej zależnej Y :

- **wzrośnie (jeżeli a>0) o $|a|$ jednostek lub**
- **spadnie (jeżeli a<0) o $|a|$ jednostek.**

b **– wyraz wolny**

INTERPRETACJA: stały poziom wartości zmiennej zależnej Y niezależny od zmian wartości zmiennej niezależnej X.

APROKSYMACJA – REGRESJA LINIOWA

Chcemy narysować prostą, która ma jak najlepiej odwzorować dany zbiór punktów (obserwacji).

Przyjmujemy równanie prostej:

$$y = a_0 + a_1 x + e \tag{1.15}$$

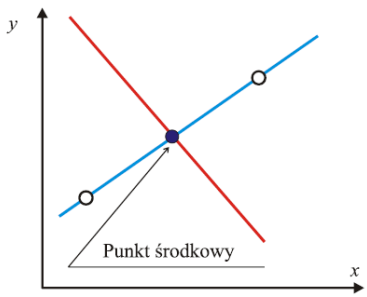
gdzie e jest błędem pomiędzy przyjętym równaniem prostej a danym punktem.

W celu dopasowania jak najlepszego równani (współczynniki a_0 i a_1) będziemy minimalizowali sumę błędów e.

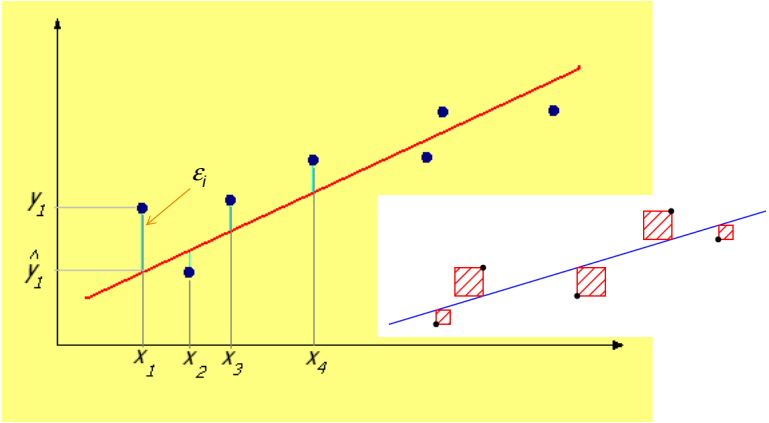
Możemy to zrobić następująco:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$

gdzie n jest liczbą punktów.



Estymacja parametrów modelu (2)



gdzie (y_i, x_i) oznacza elementy próby losowej.

Estymacja parametrów modelu (3)

Każdą obserwację empiryczną można zapisać jako:

$$y_i = b + a \cdot x_i + \varepsilon_i.$$

Problem estymacji sprowadza się do wyznaczenia miniumum funkcji *s* danej wzorem.

$$s(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - (b + a \cdot x_i)\right]^2$$

Estymacja parametrów modelu (4)

Funkcja *s* jest funkcją dwóch niewiadomych (*a* i *b*), aby znaleźć minimum tej funkcji musimy wyznaczyć pochodne cząstkowe funkcji *s* względem obu niewiadomych:

$$s'_b = -2\sum_{i=1}^{n}(y_i - b - a \cdot x_i)$$

$$s'_a = -2\sum_{i=1}^{n}x_i \cdot (y_i - b - a \cdot x_i)$$

i przyrównać te pochodne do zera.

Estymacja parametrów modelu (5)

Otrzymujemy układ równań postaci:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \hat{b} - \hat{a} \cdot x_i) = 0 \\ \sum_{i=1}^{n}x_i(y_i - \hat{b} - \hat{a} \cdot x_i) = 0 \end{cases}$$

Rozwiązując mamy:

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\operatorname{cov} xy}{\operatorname{var} x} \qquad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$$

cov *xy* (kowariancja) - liczba określająca zależność liniową między zmiennymi losowymi x i y.
var *x* (wariancja) – miara zmienności związana ze zróżnicowaniem zbioru

• Przykład 1.4

Wyznacz prostą aproksymującą dane z dwóch pierwszych kolumn tabeli

x_i	y_i
1.0	0.5
2.0	2.5
3.0	2.0
4.0	4.0
5.0	3.5
6.0	6.0
7.0	5.5

Rozwiązanie:

Obliczamy kolejno:

$n = 7, \qquad \sum x_i y_i = 119.5, \qquad \sum x_i^2 = 140$

$\sum x_i = 28, \qquad \overline{x} = \frac{28}{7} = 4$

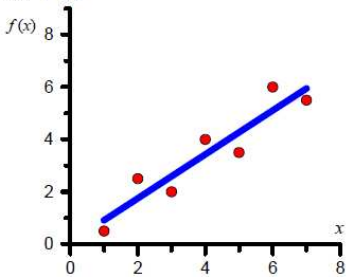
$\sum y_i = 24, \qquad \overline{y} = \frac{24}{7} = 3.428571429$

$a_1 = \frac{7(119.5) - 28(24)}{7(140) - 28^2} = 0.839285714$

$a_0 = 3.428571429 - 0.839285714(4) = 0.07142857$

Ostateczne równanie prostej uzyskane metodą najmniejszych kwadratów:

$y = 0.07142857 + 0.839285714x$



REGRESJA LINIOWA – BŁĘDY OBLICZEŃ

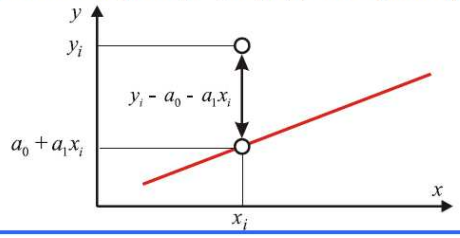
Metoda najmniejszych kwadratów wyznacza równanie prostej w sposób jednoznaczny.

Poszukujemy dodatkowych własności wyznaczonego równania.

Wracamy do równania błędów:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right)^2 \qquad (1.19)$$

Wyrażenie $(y_i - a_0 - a_1 x_i)$ jest odległością pomiędzy punktem z doświadczenia (np. pomiarem) a punktem na prostej regresji (rysunek poniżej).



Zauważmy ponadto, że:

1) rozproszenie punktów danych jest podobne w całym zakresie,
2) to rozproszenie może być opisane rozkładem normalnym

Wtedy możemy wykorzystać znane wzory z rachunku prawdopodobieństwa.

Odchylenie standardowe (standardowy błąd) linii regresji można wyznaczyć następująco

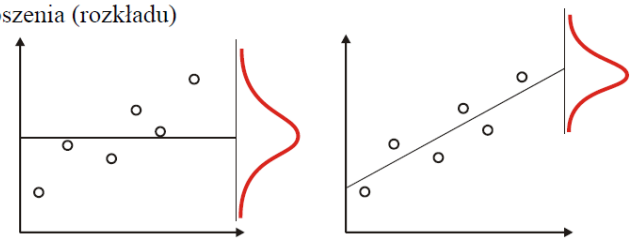
$$S_{y/x} = \sqrt{\frac{S_r}{n-2}} \qquad (1.20)$$

Oznaczenie $S_{y/x}$ oznacza, że błąd dotyczy przewidywanej wartości y dla danej wartości x.

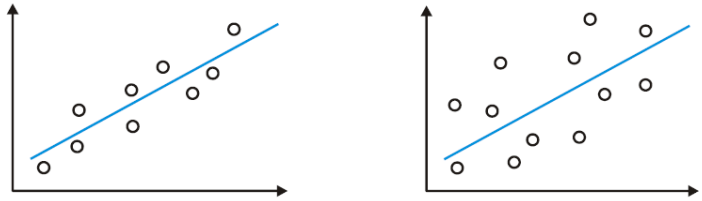
W mianowniku mamy $(n-2)$, gdyż do wyznaczenia S_r wykorzystywaliśmy dwa parametry a_0 i a_1.

Ponadto zauważmy, że nie istnieje „rozproszenie danych" dookoła prostej poprowadzonej przez dwa punkty. A więc po wstawieniu $n = 2$ mamy dzielenie przez 0.

Porównanie rozproszenia (rozkładu)



Powinniśmy także porównać jakość naszego dopasowania krzywej, które np. będzie różne dla różnych danych opisanych tą samą prostą regresji (rysunek).



Ocena dopasowania funkcji regresji do danych empirycznych

Problem oceny dopasowania był już częściowo omawiany (wykład 6) przy okazji analitycznego wygładzania szeregu czasowego za pomocą liniowej funkcji trendu.

Podstawowymi miarami „dobroci" dopasowania linii regresji do danych empirycznych są:

- **współczynnik zbieżności (φ^2)**
- **współczynnik determinacji (R^2)**
- **średni błąd szacunku (pierwiastek z tzw. wariancji resztowej)**

Współczynnik zbieżności (φ^2):

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{gdzie} \quad 0 \le \varphi^2 \le 1$$

Im φ^2 jest bliższy 0, tym dopasowanie jest lepsze.

Współczynnik determinacji (R^2):

$$R^2 = 1 - \varphi^2 \qquad \text{gdzie} \qquad 0 \le R^2 \le 1$$

Przy zależności liniowej można go wyznaczyć również jako:

$$R^2 = r_{xy}^2 \qquad \text{lub} \qquad R^2 = r_{yx}^2$$

Im R^2 jest bliższy 1, tym dopasowanie jest lepsze.

Średni błąd szacunku (S_e):

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

gdzie:

k – liczba szacowanych parametrów funkcji regresji

(tutaj k=2; szacujemy dwa parametry: a i b)

Jest to pierwiastek z wariancji resztowej (S_e^2).

Nazwa bierze się od reszty (e_i), którą definiuje się jako:
różnicę pomiędzy wartością empiryczną, a wartością teoretyczną cechy zależnej Y:

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

Pozwala na to **współczynnik determinacji**:

$$r^2 = \frac{S_t - S_r}{S_t} \tag{1.21}$$

gdzie S_t jest sumą kwadratów dookoła wartości średniej

$$S_t = \sum (y_i - \overline{y})^2, \qquad \overline{y} = \frac{\sum y_i}{n}$$

Można też wyznaczyć **współczynnik korelacji**:

$$r = \sqrt{r^2} \tag{1.22}$$

Dla $S_r = 0$, $r = r^2 = 1$ mamy 100% odwzorowanie wyników.

Alternatywny wzór pozwalający wyznaczyć **współczynnik korelacji**:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum x_i}{\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{1.23}$$

PRZYKŁAD 5

Ocena dopasowania funkcji regresji dla danych z przykładu 1.

$$\hat{y}_i = 0{,}085 \cdot x_i + 1{,}47 \qquad \overline{y} = 3$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	y_i	x_i	$\hat{y}_i$	$(y_i - \overline{y})$	$(y_i - \hat{y}_i)$	$(y_i - \overline{y})^2$	$(y_i - \hat{y}_i)^2$
1	2,0	5					
2	2,5	13					
3	2,5	16	2,83	-0,5	-0,33	0,25	0,1089
4	4,0	28	3,85	1,0	0,15	1,00	0,0225
5	5,0	42	5,04	2,0	-0,04	4,00	0,0016
6	3,0	16	2,83	0,0	0,17	0,00	0,0289
7	2,0	6	1,98	-1,0	0,02	1,00	0,0004
razem	X	X	X	X	X		

Współczynnik zbieżności

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{0{,}1787}{7{,}5} = 0{,}024$$

Współczynnik determinacji

$$R^2 = 1 - \varphi^2 = 1 - 0{,}024 = 0{,}976$$

lub wg innego wzoru

$$R^2 = r_{xy}^2 = (0{,}993)^2 = 0{,}986$$

Uwaga! Różnice w wartości współczynnika determinacji wynikają z błędów zaokrągleń na etapie liczenia współczynników: zbieżności i korelacji

Średni błąd szacunku

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{0{,}1787}{7-2}} = 0{,}189$$

W celu wyrobienia sobie poglądu nt. wielkości tego błędu odniesiemy go średniego poziomu cechy Y:

$$S_e / \bar{y} \times 100\% = 0{,}189 / 3 \times 100\% = 6{,}3\%$$

Uwaga! Nie można użyć znanego współczynnika zmienności (V_x) ponieważ średnia wartość reszt jest teoretycznie równa 0. Wystąpiłoby zatem dzielenie przez zero.

PODSUMOWANIE (przykład 5)
Wszystkie policzone miary dopasowania potwierdzają bardzo dobre dopasowanie funkcji regresji do danych empirycznych.

PRZYKŁAD 6

Na zakończenie wyznaczymy funkcję regresji dla danych z przykładu 3. Badaniu poddano tam 400 studentów. Wcześniej otrzymaliśmy tam:

$$n = 400 \quad \overline{x} = 12{,}6 \quad \overline{y} = 3{,}5 \quad s_x^2 = 47{,}04$$

$$C(X,Y) = 6{,}33 \qquad r_{xy} = 0{,}942$$

Parametry funkcji regresji wynoszą:

$$a = \frac{C(X,Y)}{s_x^2} = \frac{6{,}33}{47{,}04} = 0{,}135$$

$$b = \overline{y} - a\overline{x} = 3{,}5 - 0{,}135 \times 12{,}6 = 1{,}799$$

Funkcja regresji w przykładzie 3 ma postać:

$$\boxed{\hat{y}_i = 0{,}135 \cdot x_i + 1{,}799}$$

Dobroć dopasowania do danych empirycznych mierzona współczynnikiem determinacji wynosi:

$$R^2 = r_{xy}^2 = (0{,}942)^2 = 0{,}887$$

Powyższa funkcja regresji w 88,7% objaśnia kształtowanie się oceny z egzaminu (Y) w zależności od czasu nauki (X).

WYKORZYSTANIE funkcji regresji do PROGNOZY oceny.

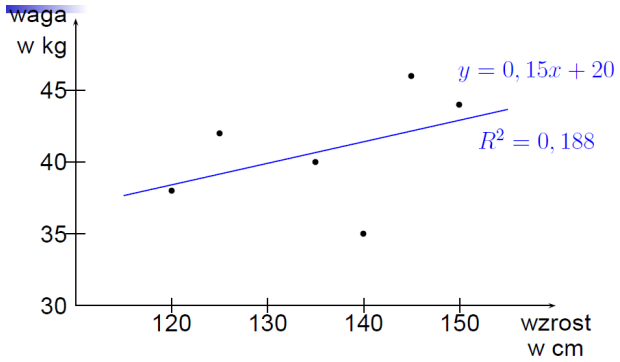
Słuchacz o numerze 401 poświęcił na naukę 20 dni (x_{401}=20).
Jakiej oceny może się spodziewać (średnio) ?

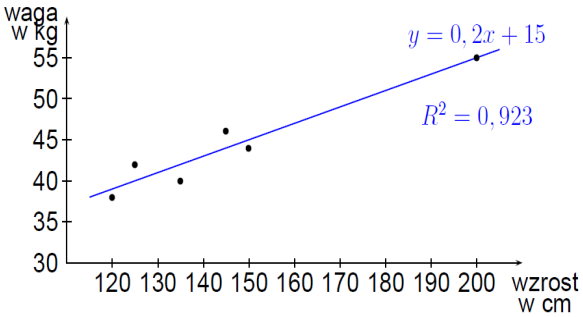
$$\hat{y}_{401} = 0{,}135 \cdot x_{401} + 1{,}7999 = 0{,}135 \times 20 + 1{,}799 = 4{,}499$$

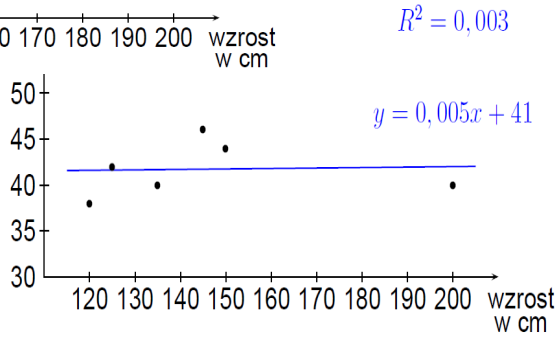
Poświęcając 20 dni na naukę słuchacz może spodziewać się (średnio !!!) oceny 4,499 czyli „db+”.

Przypadki odstające

Pojedyncze nietypowe obserwacje mogą wpływać znacząco na przebieg linii regresji, nazywa się je wtedy *obserwacjami wpływowymi*.







John Francis Anscombe (1918-2001), statystyk angielski, podał przykład 4 par zmiennych x i y, dla których otrzymujemy takie same wzory na prostą regresji i taki sam współczynnik dopasowania r^2, a tylko dla jednej z tych par model jest właściwy. Trzeba zawsze pamiętać o wykonaniu wykresów rozrzutu!

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

REGRESJA WIELOMIANOWA

Często nie jest możliwe przyjęcie prostej, jako funkcji odwzorowującej dane.
Można wtedy zastosować regresję wielomianową.

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m + e$$

Suma błędów

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m)^2 \qquad (1.24)$$

Obliczamy kolejne pochodne:

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m)$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m)$$

$$\ldots$$

$$\frac{\partial S_r}{\partial a_m} = -2\sum_{i=1}^{n} x_i^m (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m)$$

Po przyrównaniu do zera otrzymamy

$$a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 + \ldots + a_m \sum x_i^m = \sum y_i$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 + \ldots + a_m \sum x_i^{m+1} = \sum x_i y_i$$

$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 + \ldots + a_m \sum x_i^{m+2} = \sum x_i^2 y_i$$

$$\ldots$$

$$a_0 \sum x_i^m + a_1 \sum x_i^{m+1} + a_2 \sum x_i^{m+2} + \ldots + a_m \sum x_i^{2m} = \sum x_i^m y_i$$

Mamy $(m + 1)$ niewiadomych $a_0, a_1, a_2, \ldots, a_m$.
Należy rozwiązać układ $(m + 1)$ liniowych równań.
Błąd wyznaczamy z następującego wzoru

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

Gdzie m jest stopniem wielomianu.
Stopień swobody: $(m + 1)$.

• **Przykład 1.5**

Wyznacz wielomian drugiego stopnia opisujący dane z tabeli.

x_i	y_i
0.0	2.1
1.0	7.7
2.0	13.6
3.0	27.2
4.0	40.9
5.0	61.1

Rozwiązanie:

Obliczamy kolejno:

$m = 2, \qquad n = 6, \qquad \overline{x} = 2.5, \qquad \overline{y} = 25.433$

$\sum x_i = 15, \qquad \sum y_i = 152.6, \qquad \sum x_i^2 = 55, \qquad \sum x_i^3 = 225$

$\sum x_i^4 = 979, \qquad \sum x_i y_i = 585.6, \qquad \sum x_i^2 y_i = 2488.8$

Układ równań liniowych

$6a_0 + 15a_1 + 55a_2 = 152.6$

$15a_0 + 55a_1 + 255a_2 = 585.6$

$55a_0 + 225a_1 + 979a_2 = 2488.8$

Po rozwiązaniu

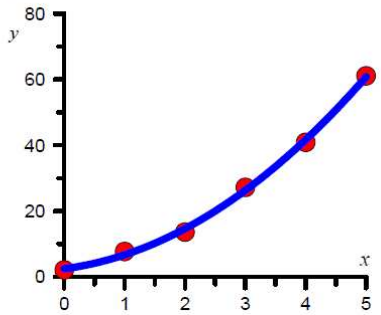
$a_0 = 2.47857$

$a_1 = 2.35929$

$a_2 = 1.86071$

Ostateczne równanie krzywej regresji:

$y = 2.47857 + 2.35929x + 1.86071x^2$



Błędy:

$$S_t = \sum_{i=1}^{6} (y_i - \bar{y})^2 = 2513.39$$

$$S_r = \sum_{i=1}^{6} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right)^2 = 3.74657$$

Odchylenie standardowe

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}} = \sqrt{\frac{3.74657}{6-(2+1)}} = 1.12$$

Współczynnik determinacji

$$r^2 = \frac{S_r - S_t}{S_t} = \frac{2513.39 - 3.74657}{2513.39} = 0.99851$$

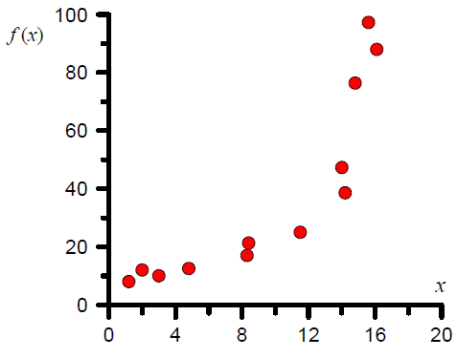
Współczynnik korelacji

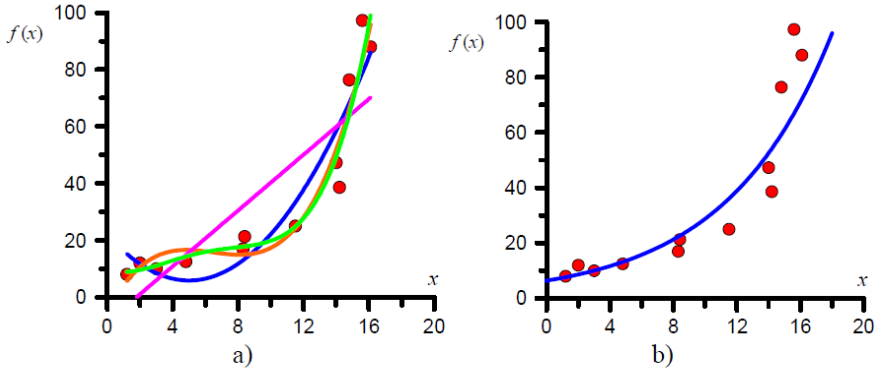
$$r = \sqrt{r^2} = \sqrt{0.99851} = 0.99925$$

- **Przykład 1.6**

Dokonaj aproksymacji mając dany zbiór danych empirycznych przedstawione w tabeli (zatrudnienie w przemyśle w zależności od dochodu).

x_i	y_i
2.0	12.0
1.2	8.0
14.8	76.4
8.3	17.0
8.4	21.3
3.0	10.0
4.8	12.5
15.6	97.3
16.1	88.0
11.5	25.0
14.2	38.6
14.0	47.3





a) Aproksymacja prostą oraz wielomianami stopnia 2 (kolor niebieski), 3 (kolor pomarańczowy) oraz 4 (kolor zielony)

b) Aproksymacja **funkcją wykładniczą** $y = ab^x$. Jak zostały obliczone parametry a i b.

Jeżeli zapiszemy logarytm funkcji wykładniczej $y = ab^x$ otrzymamy:

$$\log y = \log a + x \log b$$

czyli równanie prostej

$$Y = A + x\,B$$

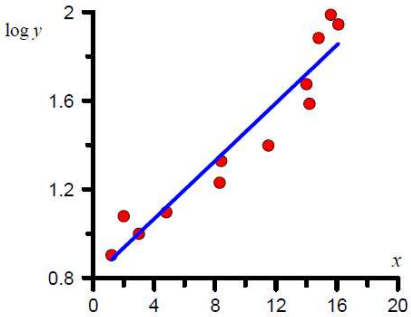
Możemy zastosować standardowe działania wyznaczając $A = \log a$ oraz $B = \log b$.

Równanie prostej

$$\log y = 0.8063 + 0.0653x$$

Należy sprawdzić czy współczynnik korelacji jest dostatecznie bliski 1.0. Będzie to świadczyć o dobrym lub złym odwzorowaniu funkcji:

$$r_{\log y, x} = 0.959$$



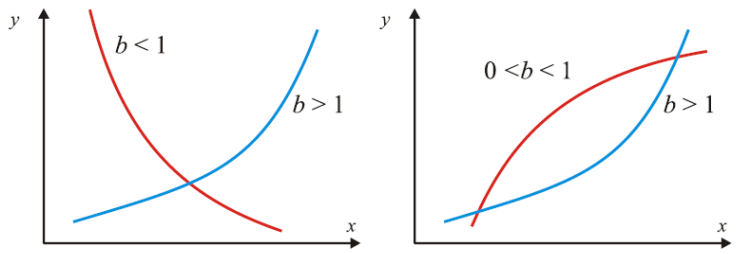
Ostatecznie funkcja aproksymująca $y = 6.401 \times 1.162^x$

UWAGA:

Od nas zależy jakie funkcje przyjmiemy jako funkcje aproksymujące nasze dane.

Rozpatrzmy jeszcze raz funkcję: $y = ax^b$

Warto jednak wiedzieć jakiego wykresu aproksymującego możemy się spodziewać dobierając odpowiednio parametry .



Linearyzacja modelu regresji wielomianowej

$$y = \frac{1}{a+bx};\ u = x, v = \frac{1}{x}$$

$$y = \frac{x}{a+bx};\ u = \frac{1}{x}, v = \frac{1}{y}$$

$$y = ax^b + c;\ u = \lg x, v = \lg(y-c)$$

$$y = ax^{\alpha};\ u = \lg x, v = \lg y$$

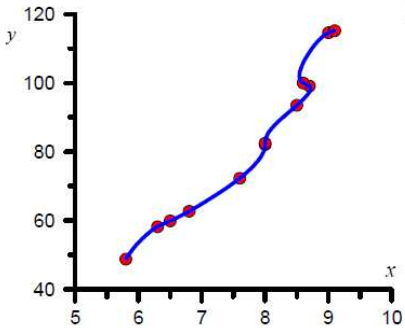
$$y = ae^{bx};\ u = x, v = \lg y$$

$$y = ae^{\frac{b}{x}};\ u = \frac{1}{x}, v = \lg y$$

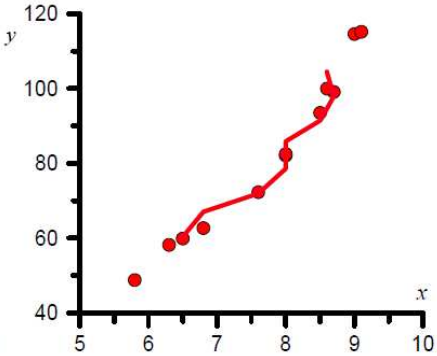
$$y = a\lg x + b;\ u = \lg x, v = y$$

Programy komercyjne

Inne możliwości dostępne w wielu komercyjnych programach, np. Statistica, Grapher, Excel, Matlab.



Spline smoothing



Running average i Weighted average

Użycie regresji:

1. **konstruowanie modelu - budowa tzw. modelu regresyjnego**, czyli funkcji, opisującej jak zależy wartość oczekiwana zmiennej objaśniającej od zmiennych objaśnianych.

Funkcja ta może być zadana:

nie tylko czystym wzorem matematycznym, ale także całym algorytmem, np. w postaci drzewa regresyjnego, sieci neuronowej, itp. Model konstruuje się tak, aby jak najlepiej pasował do danych z próby, zawierającej zarówno zmienne objaśniające, jak i objaśniane (tzw. **zbiór uczący).**

2. Wyliczanie regresji (stosowanie modelu, scoring)

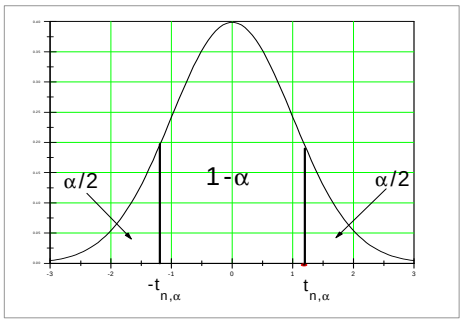
użycie wyliczonego modelu do danych w których znamy tylko zmienne objaśniające (wejściowe), w celu wyznaczenia wartości oczekiwanej zmiennej objaśnianej.

Pytanie: czy wyznaczona prosta prezentuje zależność y od x w całej populacji?
Tak, jeśli spełnione są pewne warunki. Podstawowy: błędy ε_i powinny mieć rozkład normalny o średniej 0. Drugi warunek: test badający istnienie związku liniowego pomiędzy zmiennymi powinien dawać małą istotność, tzn. mniejszą niż 0,05 czy 0,1. Test ten ma hipotezę zerową mówiącą o braku takiego związku. Dla małych wartości istotności hipotezę taką możemy odrzucić.

Należy sprawdzić

1) **Adekwatność funkcji -** czy funkcja jest odpowiednia dla badania ich zmiennych X,Y

2) **Istotność parametrów funkcji** - mówi nam, w jakim stopniu, w ilu procentach można zawierzyć ich wiarygodności (i czy w ogóle).



Poziomem istotności nazywamy przyjęte prawdopodobieństwo pomyłki w trakcie oceny istotności parametru

Werykacja adekwatnoci funkcji: **test F**

weryfikacja istotności współczynników funkcji obiektu - **test t Studenta**.

Weryfikacja adekwatności modelu obiektu

Model obiektu opisuje jego właściwości i zachowanie tylko w przybliżeniu. Spowodowane jest to niedokładnością wyznaczenia parametrów modelu oraz nieadekwatnością struktury modelu.

Na **niedokładność wyznaczenia parametrów modelu** mają wpływ następujące czynniki:
- błędy przyjętej metody identyfikacji parametrów modelu,
- błędy obliczeń numerycznych,
- błędy danych użytych do identyfikacji parametrów modelu.

Nieadekwatność struktury modelu wynika natomiast z trzech czynników:
- pominięcia wśród wielkości modelujących obiekt, czynników istotnych dla przebiegu zjawisk w obiekcie,
-niewłaściwą specyfikacją wielkości modelujących obiekt,
-przyjęciem niewłaściwego typu równania modelu.

Weryfikacja adekwatnosci modelu

Model obiektu opisuje jego właściwości i zachowanie tylko w przybliżeniu. Spowodowane jest to:

- niedokładnością wyznaczenia parametrów modelu
- nieadekwatnością struktury modelu.

Na niedokładność wyznaczenia parametrów modelu mają wpływ następujące czynniki:

- błędy przyjętej metody identyfikacji parametrów modelu,
- błędy obliczeń numerycznych,
- błędy danych użytych do identyfikacji parametrów modelu.

Weryfikacja adekwatnosci modelu

Nieadekwatność struktury modelu wynika natomiast z trzech czynników:

- pominięcia wśród wielkości modelujących obiekt, czynników istotnych dla przebiegu zjawisk w obiekcie,
- niewłaściwą specyfikacją wielkości modelujących obiekt
- przyjęciem niewłaściwego typu równania modelu.

Weryfikacja adekwatnosci modelu

Oceny adekwatności modelu dokonuje się na dwa sposoby:

1)Pierwsza metoda polega obliczeniu wartości błędu aproksymacji wybraną funkcją f i porównaniu jej z pewną arbitralnie wybraną wartością dopuszczalną e_d. *Jeśli obliczona* wartość błędu e_{max} *jest mniejsza od* e_d *wówczas uznaje się wyznaczony model za* adekwatny.

2)Zastosowanie statystycznego **testu istotności – *testu F* (*Snedecora)***

Weryfikacja adekwatnosci modelu

Definicje błędów aproksymacji:

maksymalny bezwzględny błąd aproksymacji: $\varepsilon_{\max} = \max\left|y_u - \hat{y}_u\right|$

maksymalny błąd względny: $\delta_{\max} = \max\left(\frac{\left|y_u - \hat{y}_u\right|}{y_u}\right) \cdot 100\%$

błąd średniokwadratowy: $\delta_{\max} = \sqrt{\frac{\sum_{u=1}^{n}\left(y_u - \hat{y}_u\right)}{n}}$

Testowanie hipotez statystycznych

Definicje

Hipoteza - sąd o zbiorowości generalnej (populacji) wydany na podstawie próby statystycznej.

Rodzaje hipotez - parametryczne (o wartości przeciętnej, o wskaźniku struktury, o wariancji, itp.) oraz nieparametryczne (o rozkładzie cechy, o niezależności cech X i Y, itp.).

Hipoteza zerowa (H_0) - hipoteza sprawdzana.

Hipoteza alternatywna (H_1) - hipoteza, którą jesteśmy skłonni przyjąć gdy odrzucimy hipotezę zerową (H_0).

Test statystyczny - reguła postępowania w wyniku której odrzucimy hipotezę zerową (H_0).

Rodzaje błędów w testowaniu hipotez

	przyjąć H_0	odrzucić H_0
H_0 prawdziwa	O.K. $1 - \alpha$	błąd I-rodzaju α
H_0 fałszywa	błąd II-rodzaju β	O.K. $1 - \beta$

α - jest to prawdopodobieństwo popełnienia błędu I-rodzaju i nazywane jest **poziomem istotności**. Zwykle przyjmuje się:

$\alpha=0,05$ (używane są również poziomy: 0,1; 0,02; 0,01)

Dobry test: test w którym $\alpha \approx \beta$

Testy parametryczne

Elementarnymi testami są tutaj następujące testy:

1. **Testowanie hipotezy o wartości przeciętnej (m) oraz**
2. **Testowanie hipotezy o wskaźniku struktury (p).**

W celu porównywania obu wymienionych parametrów w dwóch zbiorowościach stosuje się następujące testy (*tylko na ćwiczeniach*):

1. **Testowanie hipotezy o równości dwóch wartości przeciętnych ($m_1=m_2$) oraz**
2. **Testowanie hipotezy o równości dwóch wskaźników struktury ($p_1=p_2$).**

Testowanie hipotezy o wartości przeciętnej (m)

Założenie: Cecha ma w populacji rozkład normalny $N(m;\sigma)$.
Założenie to można weryfikować nieparametrycznymi testami zgodności (np. test zgodności chi-kwadrat).

Formułowanie hipotez

Hipoteza zerowa (H_0) jest hipotezą „o równości" i brzmi:

$$H_0: m = m_0$$

gdzie m_0 jest konkretną wartością (liczbą).
Hipoteza alternatywna (H_1) może być sformułowana trojako (najczęściej w zależności od wyniku uzyskanego w próbie):

$$H_1: m \neq m_0 \text{ (albo } H_1: m < m_0 \text{ albo też } H_1: m > m_0\text{)}$$

Wybór hipotezy alternatywnej (H_1) ma decydujące znaczenie dla sformułowania obszaru odrzucenia.

Wnioskowanie

Jeżeli wartość sprawdzianu *T* znajdzie się:

1. **w obszarze odrzucenia, to odrzucamy H_0 i przyjmujemy H_1.**
2. **poza obszarem odrzucenia, to nie mamy podstaw do odrzucenia H_0.**

UWAGA !!! Nigdy nie mówimy o przyjęciu hipotezy H_0.

Testy nieparametryczne

Omówimy tutaj dwa spośród wielu testów nieparametrycznych:

1. **test niezależności chi-kwadrat (testowanie niezależności cechy X i cechy Y) oraz**
2. **test zgodności chi-kwadrat (testowanie zgodności rozkładu badanej cechy X z wybranym rozkładem teoretycznym).**

Hipotezą statystyczną nazywamy:
każde przypuszczenie dotyczące nieznanego rozkładu badanej cechy populacji, o prawdziwości lub fałszywości którego wnioskuje się na podstawie badanej próbki.

Przy weryfikacji hipotez postępuje się w ten sposób, że oprócz weryfikowanej hipotezy zwanej hipotezą zerową wyróżnia się jeszcze inną hipotezę K, która najczęściej wynika z celu badania statystycznego, zwaną hipotezą alternatywną.

W celu weryfikacji hipotezy budujemy funkcję opartą na próbie (najlepiej próbie losowej prostej) $\delta(X_1,....,X_n)$ **zwaną statystyką testową**.

Przy pobieraniu różnych próbek, nawet o tej samej liczności n – funkcja ta przyjmuje na ogół różne wartości, z których jedne będą świadczyły o prawdziwości weryfikowanej hipotezy a inne będą ją odrzucały.

Naturalnym zatem jest podzielenie zbioru wszystkich wartości, które może przyjąć statystyka testowa na dwa dopełniające się zbiory W i W', takie że:

$\delta(X_1,....,X_n) \in W$ - hipotezę odrzucamy;
$\delta(X_1,....,X_n) \in W'$ - hipotezę przyjmujemy.

Zbiór W nazywamy zbiorem krytycznym, zaś zbiór W' zbiorem przyjęć.

Weryfikacja adekwatnosci modelu

Testu istotności – *test F* (*Snedecora):*
Weryfikacja się statystycznie hipotezy

$$H: \sigma_a^2 \Leftrightarrow \sigma^2,$$

poprzez porównanie wariancję błędów aproksymacji (*wariancję adekwatności)* σ_a^2 z wariancją niedokładności pomiarów wielkości wyjściowej σ^2.

Przyjmuje się następujące hipotezy:
1) hipoteza zerowa *H0:* $\sigma_a^2 = \sigma^2$ oznaczająca, iż model jest adekwatny,
2) hipoteza alternatywna: *H1:* $\sigma_a^2 > \sigma^2$ oznaczająca, iż model nie jest adekwatny.

Weryfikacja adekwatnosci modelu

Procedura weryfikacji statystycznej dla jednakowej liczby powtórzeń *r we wszystkich* układach planu eksperymentu jest następująca:

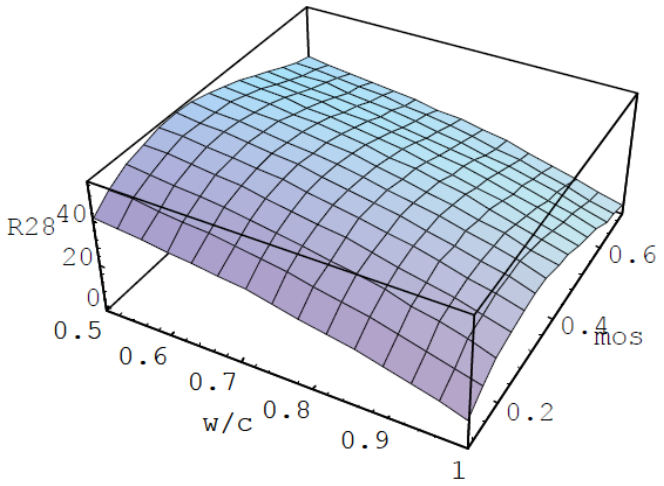
a)Oblicza się wartość funkcji testowej: $F = \frac{\sigma_a^2}{\sigma^2}$

b)na podstawie rozkładu *F (Snedecora) odczytuje się z tablicy statystycznej wartość* krytyczną F_α , f_2 , f_1 *odpowiadającą założonemu poziomowi ufności α.*

c) Sprawdza się warunek $F \leq F_\alpha, f_2, f_1$. *Jeśli warunek jest spełniony wówczas nie ma* podstaw do odrzucenia hipotezy zerowej i przyjmuje się, że model jest adekwatny. W przeciwnym razie prawdziwa jest hipoteza alternatywna, czyli model nie jest adekwatny.

Stwierdzenie na podstawie jednej z wymienionych metod nieadekwatności modelu obiektu oznacza konieczność ponownego przeprowadzenia aproksymacji funkcją o innej postaci lub zwiększenie liczby pomiarów dla każdego układu planu doświadczenia.

REGRESJA WIELOWYMIAROWA



Regresja wieloraka (1)

Dotychczas rozpatrywaliśmy tylko dwie zmienne: *Y* i *X*.

Częściej mamy do czynienia z przypadkami w których jest zmienna losową *Y* oraz *k* zmiennych *X* (stałych lub losowych).

$$y = m(x_1,...x_k) = b_0 + b_1 x_1 + \cdots + b_k x_k$$

Model regresji liniowej można również rozszerzyć w inny sposób, wprowadzając do niego jako sztucznie stworzone predyktory np. iloczyny dwóch lub większej liczby zmiennych objaśniających. Pozwala to na uwzględnienie tzw. **interakcji** pomiędzy zmiennymi, czyli zmiany siły wpływu jednej ze zmiennych przy różnych wartościach innej zmiennej.

Regresja wieloraka (2)

Współczynnik korelacji wielorakiej:

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

y_i – empiryczna wartość i-tego czynnika wynikowego,
n – ilość pomiarów,
$\bar{y}_i$ - wartość średnia i-tego czynnika wynikowego
$\hat{y}_i$ - wartość i-tego czynnika wynikowego obliczona z r. regresji

Regresja wieloraka (3)

Współczynnik R można też obliczyć na podstawie:

$$R = \sqrt{\frac{r_{y|x}^2 + r_{y|z}^2 - r_{y|x} r_{y|z} r_{x|z}}{1 - r_{y|x}^2}}$$

gdzie: $r_{y|x}$, $r_{y|z}$, $r_{x|z}$ – współczynniki korelacji liniowej pomiędzy poszczególnymi czynnikami.

Im R bliższe 1 tym wierniejsze odwzorowanie zmienności cech badanych przez liniowa funkcję regresji wielorakiej.

Regresja wieloraka (4)

Współczynniki modelu $b_1, \ldots, b_k$ będziemy nazywamy **cząstkowymi współczynnikami regresji**.

$$y_j = b_0 + b_1 x_{1j} + \cdots + b_k x_{kj} + e_j$$

Kryterium estymacji:
należy tak dobrać parametry modelu, aby suma kwadratów odchyleń od modelu była jak najmniejsza:

$$s = \sum_j e_j^2 = \sum_j \left(y_j - b_0 - b_1 x_{1j} - \cdots - b_k x_{kj} \right)^2 = \min$$

Badanie istotności regresji wielokrotnej

Hipotezę o nieistotności regresji wielokrotnej możemy zapisać jako:

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

jej weryfikacja testem *F* Fishera-Snedecora.

Sumy kwadratów odchyleń i średnie kwadraty potrzebne do zweryfikowania hipotezy o istotności regresji mogą być wyznaczone z niżej podanych wzorów:

$$SS_R = \sum_i \hat{b}_i \operatorname{cov} x_i y \qquad MS_R = \frac{SS_R}{k}$$

$$SS_E = \operatorname{var} y - \sum_i \hat{b}_i \operatorname{cov} x_i y \qquad MS_E = \frac{SS_E}{n-k-1}$$

Badanie istotności regresji wielokrotnej

Hipotezę

$$H_0: b_1 = b_2 = \cdots = b_k = 0$$

odrzucamy gdy

$$F_R > F_{\alpha,k,n-k-1}$$

Odrzucenie hipotezy H_0 jest równoznaczne z tym, że
co najmniej jeden współczynnik regresji jest różny od zera;

tzn. istnieje związek funkcyjny liniowy między zmienną zależną a zmiennymi niezależnymi.

Problem statystyczny:
które zmienne niezależne powinny pozostać w modelu regresji.

Weryfikacja istotności współczynników regresji (2)

Charakteryzując obiekt badań przyjmuje się określoną liczbę zmiennych wejściowych. Nie ma jednak pewności czy wszystkie zdefiniowane zmienne wejściowe mają wpływ na działanie obiektu.

Stwierdzenie braku skorelowania określonej zmiennej wejściowej x_k *ze* zmienną wyjściową *y umożliwia uproszczenie modelu badań poprzez usunięcie zmiennej* x_k.

Działanie takie jest uzasadnione głównie ze względów ekonomicznych, gdyż prostszy model oznacza mniejszą ilość sprzętu technicznego nie-zbędnego do przeprowadzenia pomiarów oraz uproszczenie obliczeń matematycznych.

Weryfikacja istotności współczynników regresji (2)

Informacja o wpływie kolejnych wielkości wejściowych x_k *na wielkość wyjściową y jest* ukryta w wartościach współczynników funkcji aproksymującej. Przykładowo:

jeśli wszystkie współczynniki przy x_2 *wynoszą zero tzn.* $a_2 = a_{22} = a_{12} = 0$ *wówczas można* stwierdzić, że wielkość wyjściowa *y nie zależy od wielkości wejściowej* x_2.

Gdyby natomiast współczynniki przy x_2 *wynosiły:* $a_{22} = a_{12} = 0$ *oraz* $a_2 \neq 0$ *wówczas można* wyciągnąć wniosek, że wielkość x_2 *wpływa na wielkość wyjściową, ale tylko liniowo.*

Weryfikacja istotności współczynników regresji (2)

Analiza współczynników funkcji aproksymującej jest bardzo istotna dla realizatora badań, który uzyskuje w ten sposób istotne informacje o sposobie działania obiektu.

Analiza ta nosi nazwę ***weryfikacji istotności współczynników funkcji aproksymującej.*** Realizowana jest w oparciu o *test t-Studenta oraz ocenę wartości kowariancji wszystkich* par współczynników $\{a_i, a_j\}$ *funkcji aproksymującej* $f(\cdot)$.

Wykrycie nieistotnych współczynników funkcji aproksymującej na podstawie testu t-Studenta lub ich wzajemnego skorelowania (niezerowej wartości kowariancji) wskazuje na konieczność uproszczenia modelu. Po wyznaczeniu funkcji aproksymującej należy ponownie przeprowadzić weryfikację jej adekwatności. Dopiero pozytywne przejście tej weryfikacji jest podstawą eliminacji nieistotnych współczynników.

Weryfikacja hipotez o istotności cząstkowych współczynników regresji

Problem sprowadza się do zweryfikowania serii *k* hipotez zerowych mówiących o tym, że *i*-ty cząstkowy współczynnik regresji jest równy zero.

Hipotezy te mogą być weryfikowane testem *t*-Studenta

Weryfikacja hipotez

$$H_0 : b_i = 0$$

Wyrażenie

$$s^2_{y/x_1,\ldots x_k} = \frac{\operatorname{var} y - \sum_i \hat{b}_i \operatorname{cov} x_i y}{n-k-1}$$

jest oszacowaniem średniego kwadratu odchyleń od regresji. Przy prawdziwości hipotez zerowych tak określone statystyki mają rozkład *t*-Studenta z liczbą stopni swobody równą *n-k-1*

Hipotezę

$$H_0 : b_i = 0$$

będziemy odrzucać, jeżeli wartość statystyki *t* znajdzie się w obszarze krytycznym.

Jeżeli zmienne niezależne są z sobą powiązane
to oceny istotności cząstkowych współczynników regresji nie są niezależne.

Problem doboru zmiennych

W przypadku istnienia **silnych współzależności** między zmiennymi niezależnymi analizując funkcję regresji wielokrotnej dochodzimy do wniosku, że **jest ona istotna statystycznie** (testem *F*).

Weryfikując dalej hipotezy o istotności cząstkowych współczynników uzyskujemy wartości testu *t* Studenta, które **nie przeczą hipotezom zerowym**.

Czyli mamy istotną funkcję regresji ale wszystkie zmienne (analizowane oddzielnie) są nieistotne, powinny więc być usunięte z modelu.

Zaczynamy od pełnego zestawu potencjalnych zmiennych niezależnych, a następnie kolejno usuwamy z modelu tę zmienną niezależną, której rola w opisywaniu zależności między zmienną *Y* a zmiennymi niezależnymi jest najmniejsza. Podejście takie nosi nazwę **regresji krokowej**.

Regresja nieliniowa (1)

W wielu przypadkach interesuje nas nieliniowy związek między zmienną Y a zmienną X

Właściwie *Estymację nieliniową* możemy traktować jako *uogólnienie* metod liniowych.

W przypadku *Estymacji nieliniowej* sami decydujemy o określeniu natury tego związku; na przykład możemy przyjąć, że zmienna zależna ma być funkcją:

- logarytmiczną zmiennej niezależnej (zmiennych niezależnych)
- funkcją wykładniczą
- funkcją pewnego założonego ilorazu zmiennych niezależnych itd.

Regresa nieliniowa (2)

Współczynniki regresji:
i-ty, cząstkowy współczynnik regresji opisuje o ile średnio zmieni się wartość zmiennej *Y* przy wzroście *i*-tej wartości zmiennej *X* o jednostkę przy ustalonych wartościach pozostałych zmiennych niezależnych.

W przypadku większości modeli regresji nieliniowej taka interpretracja nie jest możliwa.

Jeśli dopuszczamy dowolny typ zależności między zmiennymi niezależnymi a zmienną zależną, pojawiają się dwa pytania, po pierwsze, jakie rodzaje zależności "mają sens", to znaczy, jak można je w znaczący sposób zinterpretować?

Zależność nieliniowa nie daje się zwykle tak łatwo zinterpretować i zwerbalizować. Po drugie, jak dokładnie obliczyć zależność, to znaczy jak wywnioskować, czy faktycznie występuje zależność nieliniowa taka, jakiej oczekiwaliśmy?

Wykorzystanie funkcji regresji do prognozowania

Słuchacz o numerze 8 (przypomnijmy, że badanie przeprowadzono dla n=7 studentów) poświęcił na naukę 20 dni (x_8=20).
Jakiej oceny może spodziewać się (średnio) przy takim nakładzie czasu na naukę ?

$$\hat{y}_8 = 0{,}085 \cdot x_8 + 1{,}47 = 0{,}085 \times 20 + 1{,}47 = 3{,}17$$

Poświęcając 20 dni na naukę słuchacz może spodziewać się (średnio !!!) oceny 3,17 czyli „dst+”.

Współczynnik determinacji

$$r_{y|x}{}^2 = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}$$

Wyrażenie to nazywamy **współczynnikiem determinacji**.

$r^2 \in <0; 1>$

Informuje:

jaka część zmienności całkowitej zmiennej losowej *Y* została wyjaśniona regresją liniową względem *X*.

Weryfikacja hipotezy o istotności korelacji

Założymy, że rozkład zmiennych losowych *Y* i *X* w populacji generalnej jest normalny. Na podstawie *n* - elementowej próby chcemy zweryfikować hipotezę, że zmienne te są liniowo niezależne:

$$H_0: \rho = 0 \qquad \text{wobec} \qquad H_1: \rho \neq 0$$

Jeżeli H_0 jest prawdziwa, to statystyka:

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

ma rozkład *t* Studenta z liczbą stopni swobody $v = n - 2$.

Hipoteza o istotności korelacji może być także zweryfikowana poprzez porównanie wyznaczonego współczynnika z próby z wartościami **krytycznymi współczynnika korelacji wielokrotnej Pearsona**.

$$|r| > R_{\alpha,k,n-k-1}$$

Weryfikacja hipotezy o istotności regresji

Weryfikacji hipotezy o istotności regresji testem *F* Fishera-Snedecora.

Analiza wariancji ma postać

Zmienność	*df*	*SS*	*M.S*	$F_{emp.}$	F_{α}
Regresji	1	$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MS_R	F_R	
Odchyleń	n-2	SS_E	MS_E	$F_{\alpha,1,n-2}$	
Całkowita	n-1	$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \operatorname{var} y$			

Przedział ufności dla wartości modelu

Dla regresji liniowej statystyka:

$$t = \frac{\hat{m}(x) - m(x)}{S_{\hat{m}(x)}}$$

ma rozkład *t* Studenta z liczbą stopni swobody *n* - 2.

Na tej podstawie możemy wyznaczyć przedział ufności dla wartości z modelu:

$$m(x) \in < \hat{m}(x) - t_{\alpha,n-2} S_{\hat{m}(x)}; \hat{m}(x) + t_{\alpha,n-2} S_{\hat{m}(x)} >$$