Liniowe relacje między zmiennymi

Marta Zalewska
Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii

Ocena liniowych relacji między zmiennymi

- Metoda korelacji - określenie rodzaju i siły zależności między cechami.
- Metoda regresji

Uwaga

- Liczbowe stwierdzenie występowania zależności między x i Y nie musi oznaczać występowania zależności przyczynowo-skutkowej.
- Współzależność dwóch zmiennych może wynikać z bezpośredniego oddziaływania na nie trzeciej zmiennej, której nie rozważamy w analizowanym badaniu.

Współczynniki korelacji

- Dla cech jakościowych (bądź ilościowych skategoryzowanych) stosujemy współczynnik korelacji Spearmana.
- Dla zmiennych ilościowych ciągłych stosujemy współczynnik korelacji Pearsona.

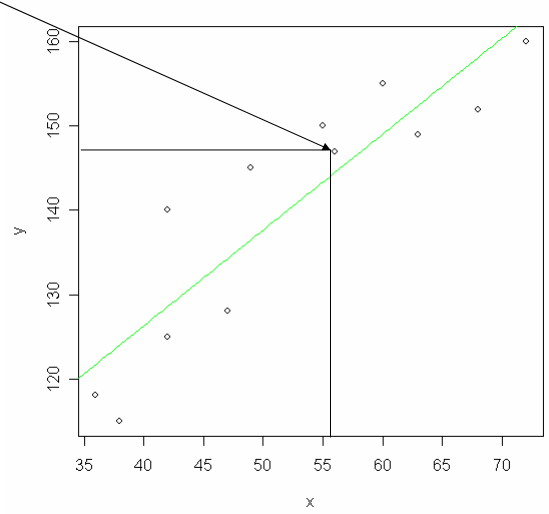
Badanie zależności między dwiema cechami

Jak zmienia się wartość jednej cechy w zależności od zmian wartości drugiej?

Wykres punktów empirycznych, diagram korelacyjny

Wykres składa się z tylu punktów ile jest obiektów w próbce. Dwa obiekty mające tę samą wartość jednej cechy mogą mieć różne wartości drugiej cechy – rozproszenie wykresu

x=c(56,42,72,36,63,47,55,49,38,42,68,60)
y=c(147,125,160,118,149,128,150,145,115,140,152,155)



Współczynnik korelacji Pearsona

- Jest miarą współzależności liniowej między dwiema **cechami ciągłymi**

  Pozwala ocenić w jakim stopniu wykres punktów indywidualnych jest bliski pewnej prostej lub czy zmiana jednej cechy powoduje proporcjonalną zmianę wartości drugiej cechy.

(najczęściej oznaczany dla próbki **r** dla populacji ρ)

Dane empiryczne dla obliczania **r:**
dany jest zbiór dwucechowych obserwacji (x_i,y_i), (i=1,2,…,n) dokonanych na na n obiektach próbki gdzie x_i,y_i oznaczają wartości cechy X i Y zaobserwowane na i-tym obiekcie.

Wzory dla obliczania **r**

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$r = \frac{\mathrm{COV}_{xy}}{S_x S_y}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Interpretacja współczynnika korelacji Pearsona (r)

- Wartość **r** zawiera się zawsze <-1,+1>,
- Wartość **r** pozwala ocenić kierunek i siłę współzależności liniowej między dwiema cechami,
- Kierunek współzależności:
- **r>0** większej wartości jednej cechy odpowiada większa wartość drugiej. Mówi się, że cechy korelują dodatnio.
- **r<0** większej wartości jednej cechy odpowiada mniejsza wartość drugiej. Mówi się, że cechy korelują ujemnie.

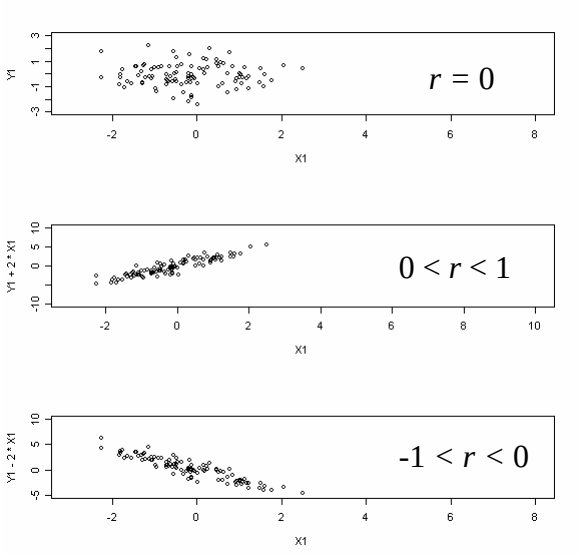
Jeżeli rozproszenie punktów jest jednakowe we wszystkich kierunkach to wartość cechy Y nie zależy od wartości cechy X

Zależność jest tym większa im bliżej pewnej prostej ciągłej znajdują się te punkty.

Prostą znajduje się najczęściej metodą najmniejszych kwadratów.

Uzyskany wzór to model matematyczny zależności

Graficzna prezentacja różnych wartości współczynnika korelacji



Współczynniki korelacji rang

- r Spearmana
- τ Kendalla

Dla cech mierzonych na skali porządkowej

Najczęściej stosowanym współczynnikiem korelacji rang jest współczynnik Spearmana.

- Jest miarą współzależności między dwiema cechami, których wartości są rangami (pozycjami) obiektów uporządkowanych osobno według jednej cechy i osobno według drugiej.

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2 - 1)}$$

Analiza regresji

- Analiza regresji zajmuje się badaniem zależności jednej cechy od innej obserwowanej cechy (cech).
- Podstawą analizowania jest model (równanie) tej zależności – **równanie regresji.**
- Równanie (wzór) wykorzystuje się do przewidywania (prognozowania, predykcji) wartości jednej cechy na podstawie wartości innej (innych) cech.

Prosta regresja liniowa

- Interesuje nas związek między dwiema zmiennymi (cechami) x i Y. Mierzymy lub obserwujemy wielokrotnie wartości tych zmiennych.
- Dane empiryczne są postaci (x_i, Y_i) - co oznacza wartości cech x i Y dla i-tego spośród n obiektów próbki.

Postać danych

Przypadki (obiekty)	Zmienna niezależna (objaśniająca) x	Zmienna zależna (objaśniana) Y
1	x_1	Y_1
2	x_2	Y_2
:	:	:
n	x_n	Y_n

Analiza zależności między zmiennymi ilościowymi

Badamy zależności między:

- dawkami pewnego preparatu a procentową zawartościa pewnego składnika krwi;
- czasem leczenia chorych a aktywnością pewnego enzymu;
- wagą a wzrostem chorych na pewną chorobę;

Przykłady zależności:

- masy mózgu człowieka i masą jego ciała;
- objętości płuc ssaków od masy ich ciała;
- liczby krwinek czerwonych a ich objętością;
- kosztami utrzymania placówki zdrowia od liczby personelu i liczby pacjentów.

Model liniowy

- Zmienna Y jest funkcją x ale zaburzoną błędami losowymi. Nasz model dla najprostszej liniowej postaci funkcji:

$$Y = a + bx + e$$

- Gdzie e jest błędem losowym o wartości oczekiwanej 0 i wariancji sigma^2. Prostą

$$y = a + bx$$

nazywamy prostą regresji

W równaniu regresji

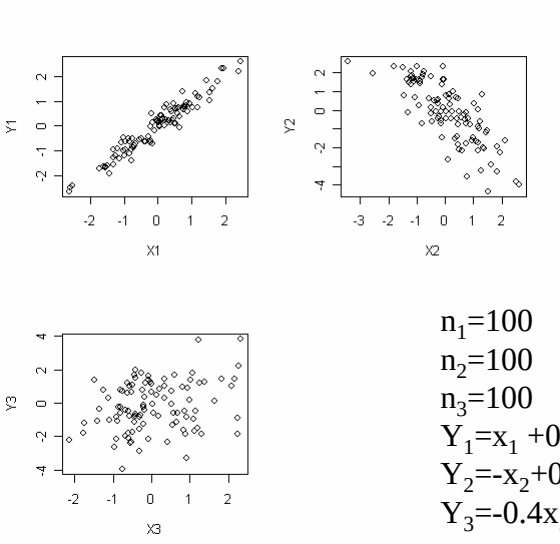
- Y- zmienna objaśniana (kryterialna, zależna).
- x - zmienna objaśniająca niezależna
- Dla poszczególnych przypadków czyli uzyskanych doświadczalnie punktów mamy (model):

$$Y_i = a + bx_i + e_i, \quad i = 1,...,n$$

Współczynniki a i b są nieznane

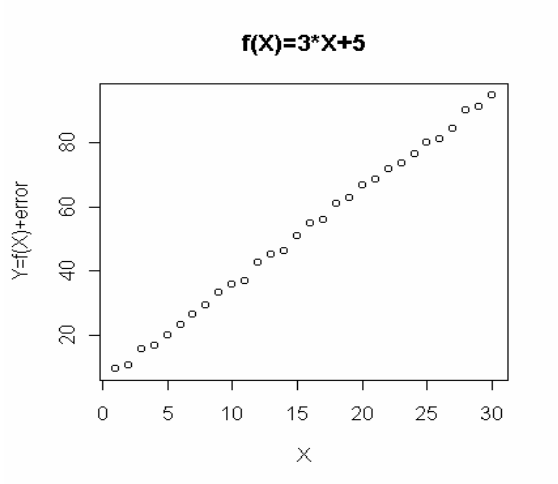
$$Y_i = a + bx_i + e_i, \quad i = 1,\ldots,n$$

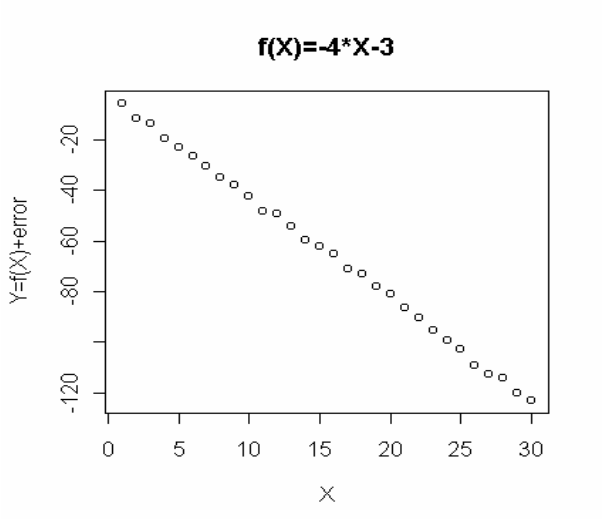
- Współczynniki a i b odgrywają rolę parametrów modelu i będziemy je estymować na podstawie danych.
- Zakładamy, że wielkości x są znane i nielosowe.
- Zmienna x jest pod kontrolą obserwatora i jest mierzona bezbłędnie.
- Wartości zmiennej Y są losowymi obserwacjami (ze względu na wpływ losowego składnika e)

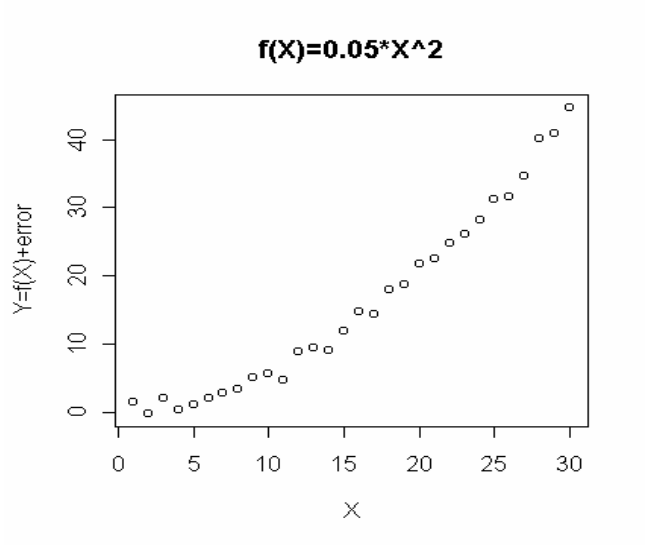


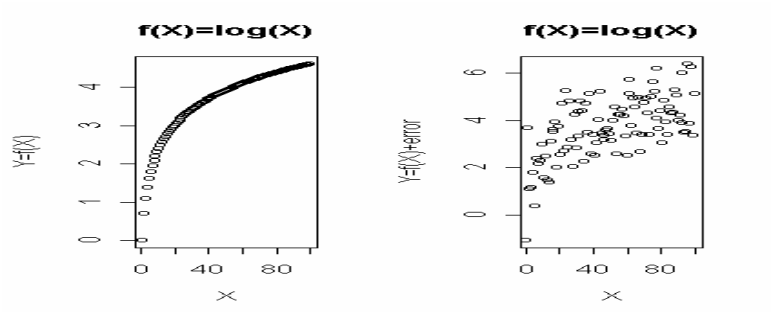
$n_1=100$
$n_2=100$
$n_3=100$
$Y_1=x_1+0.3$
$Y_2=-x_2+0.9$
$Y_3=-0.4x_3+1.5$

Przykłady przybliżonych zależności funkcyjnych






f(X)=0.05*X^2
Y=f(X)+error
0
10
20
30
40
0
5
10
15
20
25
30
X


f(X)=log(X)
Y=f(X)
0
1
2
3
4
0
40
80
X
f(X)=log(X)
Y=f(X)+error
0
2
4
6
0
40
80
X

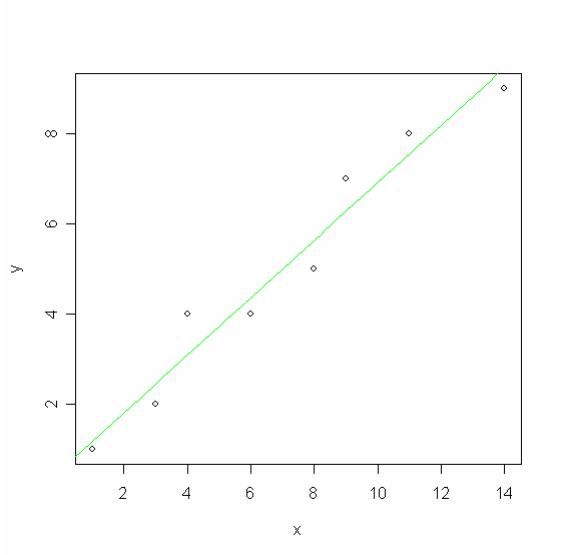
Prosta regresji

- Jest przybliżeniem wykresu punktów indywidualnych uzyskanym wg metody najmniejszych kwadratów. Równanie prostej regresji jest wzorem (modelem) na zależność liniową między dwiema badanymi cechami

Prosta regresja liniowa

- Problem badawczy: podać wzór na zależność liniową Y od x

$$y = a + bx$$



Estymatory współczynników regresji a i b z próbki

- Otrzymuje się je metodą najmniejszych kwadratów tzn. poszukując a i b takich by

$$SSE = \sum (Y_i - a - bx_i)^2 = \min$$

Prosta regresji z próbki

Minimalizując sumę kwadratów błędów
(SSE –Sum of Squares of Errors), obliczając pochodne względem a i b oraz przyrównując je do zera otrzymujemy tzw. równanie normalne, których rozwiązania są:

$$\hat{b} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} \qquad \hat{b} = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x}) x_i}$$

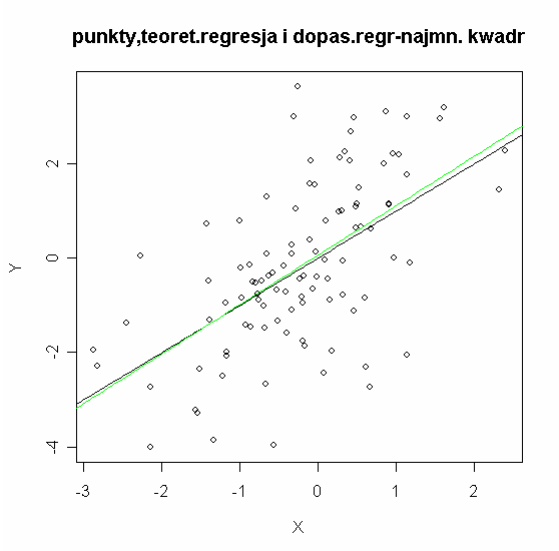
$$\hat{a} = \bar{Y} - \hat{b}\bar{x}$$

$$\bar{Y} = \frac{1}{n}\sum Y_i \qquad \bar{x} = \frac{1}{n}\sum x_i$$

Cd. Wzory na estymowane współczynniki regresji w MNK

$$\hat{b} = r\frac{S_Y}{S_X}, \hat{a} = \bar{y} - b\bar{x}$$

n – obiektów
r współ korelacji Pearsona
S_x, S_y *odchylenia st.*



Przewidywane wartości zmiennej zależnej:

$$\hat{Y}_i = \hat{a} + \hat{b}x_i$$

są to współrzędne punktów leżące na estymowanej prostej regresji

Odchylenia wielkości obserwowanych od wielkości przewidywanych nazywamy resztami:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

Reszty nie są tym samym co błędy

$\sum \hat{e}^2{}_i$ = min SSE

Estymator wariancji błędu trzeba podzielić $\sum \hat{e}^2{}_i$ przez n-2

Podstawowa tożsamość analizy wariancji

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

Całkowita zmienność Y = zmienność wyjaśniona regresją (punkty na prostej) + zmienność resztowa (albo z błędów)

To samo co r^2 $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 jest częścią zmienności wyjaśnioną przez regresję

Rozkład całkowitej zmienności Y

SST = *SSE* + *SSR*, gdzie

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Nazewnictwo:

- *SST* = total sum of squares
- *SSE* = error sum of squares
- *SSR* = regression sum of squares

Współczynnik dopasowania:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 jest częścią zmienności wyjaśnioną przez regresję

Kwadrat współczynnika korelacji r jest współczynnikiem dopasowania

$$r = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (Y_i - \bar{Y})^2}}$$

Współczynnik determinacji (dopasowania) cd.

- zmienność wyjaśniona przez model / zmienność całkowita
- określa on stopień, w jakim zależność liniowa między Y i x tłumaczy zmienność wykresu rozproszenia.
- $0 < R^2 < 1$

- Wyniki estymacji współczynników równania regresji w pakiecie statystycznym R

Rozwiązywanie zadań z regresji

- Wykres rozproszenia danych
- plot(x,y)
- Obliczanie próbkowego współczynnika korelacji Pearsona (różne możliwości) np:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

cd. rozwiązywanie zadań

- Dopasowanie prostej regresji $y=a+bx$ metodą najmniejszych kwadratów

$$\hat{b} = r\frac{S_Y}{S_X}, \hat{a} = \bar{y} - \hat{b}\bar{x}$$

- Na rysunku rozproszenia danych narysować prostą regresji
- plot(x,y)
- abline(a,b)

cd. rozwiązywanie zadań

- Oblicz współczynnik determinacji i oceń jakość dopasowania prostej regresji
- *R2*=1-*SSE*/*SST*
- $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

yhat=a+b*x (a i b z równania regresji)

cd. obliczanie współczynnika determinacji

- SSE=sum((y-yhat)^2)
- SST=sum((y-mean(y))^2)
- R2=1-SSE/SST
- R2

cd zadania z regresji - użycie funkcji lm

- z=lm(y~x)
- summary(z) # podsumowanie wyników analizy regresji

cd. predykcja

- Do wzoru na regresję liniową wstawiamy ten x dla którego dokonujemy predykcji $Y^*=a+bx$ i obliczamy Y^*

Wyniki estymacji wsp. regresji w pakiecie statystycznym R. Przykład 1

Residuals:

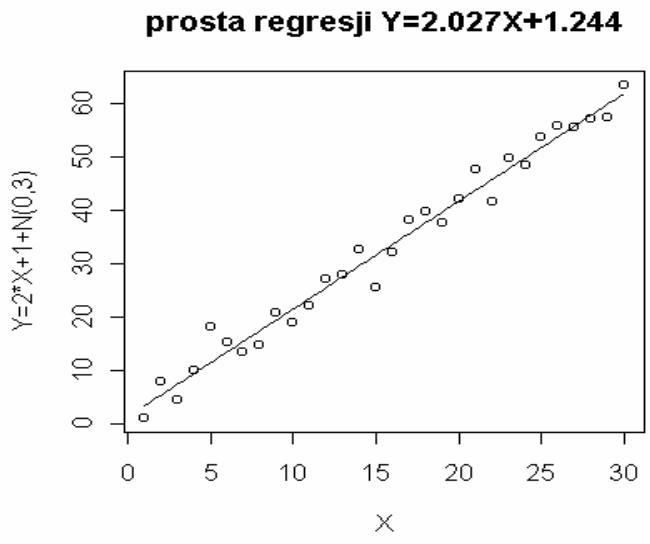
Min	1Q	Median	3Q	Max
-6.217	-2.114	0.289	1.885	6.826

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2442	1.0404	1.196	0.242
x	2.0272	0.0586	34.594	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.778 on 28 degrees of freedom
Multiple R-Squared: 0.9771, Adjusted R-squared: 0.9763
F-statistic: 1197 on 1 and 28 DF, p-value: < 2.2e-16



Wnioski z przykładu 1:

- Równanie regresji : $Y = 2.0272X + 1.2442$
- wsp. kierunkowy regresji jest istotny na poziomie istotności $< 2*10^{-16}$
- wyraz wolny jest istotny na poziomie istotności 0,20
- prosta jest dobrze dopasowana do danych, bo R^2 wynosi 0,9771

Przykład 2. Residuals:

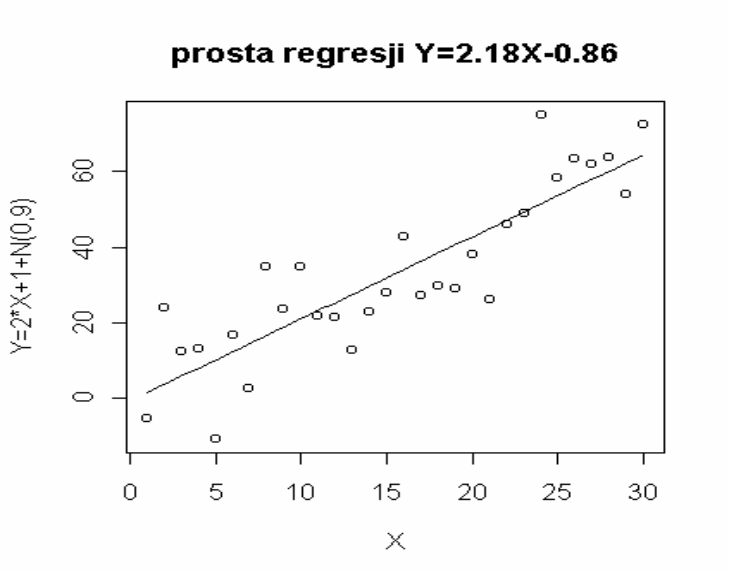
Min	1Q	Median	3Q	Max
-21.0779	-8.0028	-0.7656	6.2725	23.6319

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.8639	4.2157	-0.205	0.839	
x	2.1813	0.2375	9.186	6.06e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.26 on 28 degrees of freedom
Multiple R-Squared: 0.7508, Adjusted R-squared: 0.7419
F-statistic: 84.38 on 1 and 28 DF, p-value: 6.064e-10



Wnioski z przykładu 2:

- Równanie regresji : $Y = 2.18x - 0.86$
- współczynnik kierunkowy regresji (x) jest istotny na poziomie istotności $< 6*10^{-10}$
- wyraz wolny (Intercept) jest nieistotny
- prosta jest gorzej dopasowana do danych niż w poprzednim przykładzie, bo R^2 spadł z 0,9771 do 0,75.

Przykład 3. Wyniki regresji:

Residuals:

Min	1Q	Median	3Q	Max
-99.652	-22.389	6.736	27.834	101.244

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.093	15.890	1.139	0.265
x	1.081	0.895	1.207	0.237

Parametry nieistotne

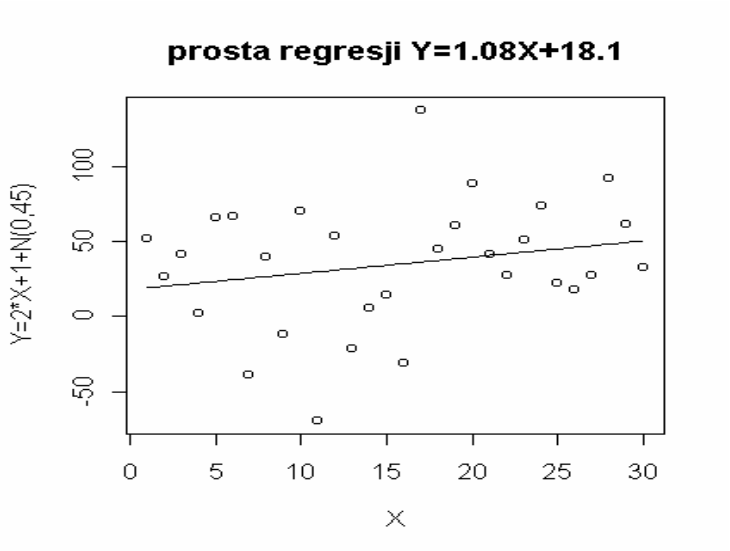
Residual standard error: 42.43 on 28 degrees of freedom

Multiple R-Squared: 0.04948, **mała wartość wsp. determinacji**

Adjusted R-squared: 0.01554

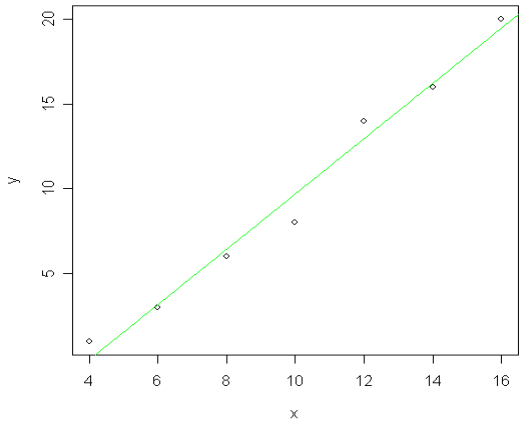
F-statistic: 1.458 on 1 and 28 DF, p-value: 0.2374

nieistotny związek regresyjny

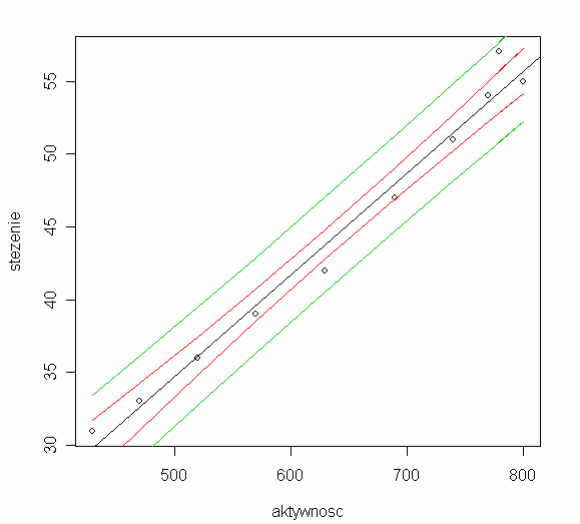


Przykład prognozy

Na podstawie danych metodą najmniejszych kwadratów otrzymano prostą regresji: Y= X-2, prognozą dla X = 7 jest Y* = 5



Zbiory ufności dla prostej regresji



Model regresji liniowej wielozmiennej

- $Y = a_0 + a_1X_1 + ...+ a_kX_k + error$, gdzie
- Y - zmienna objaśniana (typu ciągłego)
- $X_1, ... ,X_k$ zmienne objaśniające (typu ciagłego)
- $a_0, a_1, ... , a_k$ - parametry modelu
- *error* - błąd losowy