Matematyczne aspekty analizy danych
semestr zimowy 2024/2025

Dr Anna Muranova
UWM w Olsztynie

Wykład 9

Algorytm: wektory własne

Szukamy λ taki ze $|A - \lambda I| = 0$. Szukamy wektory własne $|A - \lambda I|\overline{x} = \overline{0}$.

Znaleźć wartości i wektory własne macierzy

$$\begin{pmatrix} 5 & 4 & 4 \\ -3 & -3 & -5 \\ 1 & 2 & 4 \end{pmatrix}.$$

(na tablice)

```
import sympy as sp

A = sp.Matrix([[5,4, 4],[-3, -3, -5],[1,2, 4]])
x, y,z = sp.symbols('x, y,z')
X = sp.Matrix([[x],[y],[z]])
for i in 1,2,3:
    print(sp.solve ([(A-i*sp.eye(3))*X],x,y,z,dict=True))
```

Python

```
import numpy as np

A = np.array([[5,4, 4],[-3, -3, -5],[1,2, 4]])
print(np.linalg.eig(A))
```

```
import sympy as sp

A = sp.Matrix([[5,4, 4],[-3, -3, -5],[1,2, 4]])
print(A.eigenvects())
#(eigenvalue, algebraic_multiplicity, [eigenvectors])
```

Czy zawsze istnieją wektory własne?

Macierz może nie mieć wektorów własnych. Rozważmy macierz

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Aby znaleźć wartości własne, obliczamy wyznacznik:

$$0 = \begin{vmatrix} -\lambda & -1 \\ 1 & -\lambda \end{vmatrix} = \lambda^2 + 1.$$

Rozwiązań możemy jednak szukać wśród liczb zespolonych:

$$\lambda = \pm i, \quad v = \begin{pmatrix} \pm i \\ 1 \end{pmatrix}.$$

Dopuszczając liczby zespolone, zawsze będziemy mieli co najmniej jeden wektor własny.

Uwagi

- W zastosowaniach wygodnie jest mieć bazę złożoną z wektorów własnych rozważanej macierzy.
- Wektory własne należące do różnych wartości własnych są liniowo niezależne. Dlatego, gdy wszystkie wartości własne są różne, istnieje baza złożona z wektorów własnych.
- W przypadku macierzy symetrycznej zawsze znajdziemy bazę złożoną z wektorów własnych.
- Czasem taka baza nie istnieje. Ale dla większości macierze – istnieje ($\Leftrightarrow$ macierz jest diagonalizowalna $\Leftrightarrow$ $A = BDB^{-1}$).

Rozważmy macierz $M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Dla tej macierzy mamy $M^2 = 0$, dlatego wartości własne mogą być tylko zerami (można to sprawdzić metodą zastosowaną w zadaniach). Gdyby wektory własne u i v tworzyły bazę, to dla dowolnego wektora w mielibyśmy:

$$Mw = M(au + bv) = aMu + bMv = 0.$$

Oznaczałoby to, że $M = 0$, co jest sprzeczne z definicją macierzy M.

Rozdział 6. Regresja liniowa

Regresja liniowa

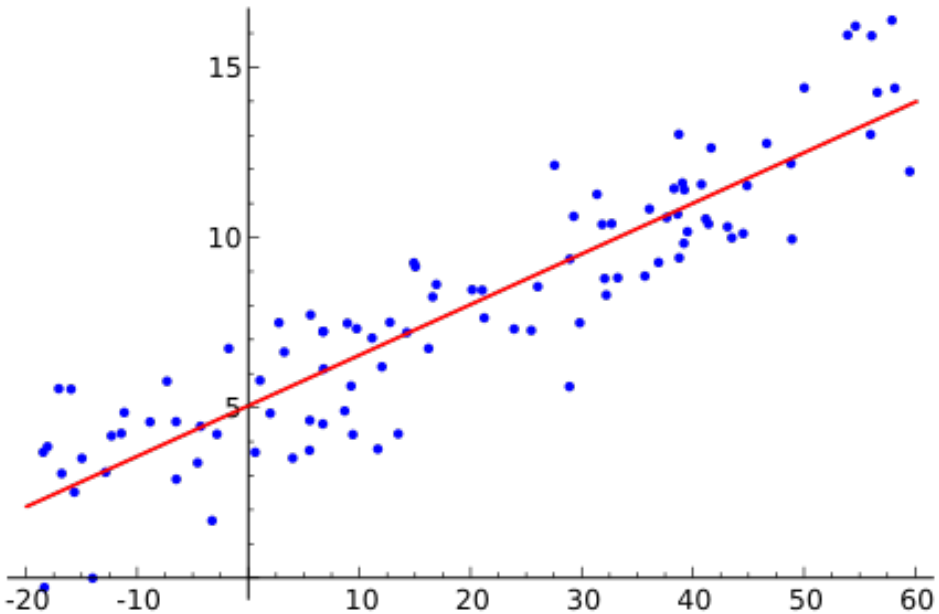
Niech mamy dwie cechy x i y. Regresja jest metodą budowania krzywej zależności Y od X, na podstawie ich wartości w probie, $X_1, \ldots, X_n$ oraz $Y_1, \ldots, Y_n$. Założenie modelu:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$$

gdzie β_1, β_2 są parametrami, a ϵ_i – wartość losowa błędu. Regresja liniowa – metody oparte o liniowe kombinacje zmiennych i parametrów dopasowujących model do danych. Dopasowana linia lub krzywa regresji reprezentuje oszacowaną wartość oczekiwaną zmiennej y przy konkretnych wartościach innej zmiennej lub zmiennych x. W najprostszym przypadku dopasowana jest stała lub funkcja liniowa, tzn. że regresja liniowa jest metodą wyznaczenia parametrów najlepiej dopasowanej prostej

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i, \quad \hat{\beta}_1 = ?, \hat{\beta}_2 = ?$$

Regresja liniowa



Rysunek: źródło: `https://en.wikipedia.org/wiki/Linear_regression`

Regresja liniowa

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i, \quad \hat{\beta}_1 = ?, \hat{\beta}_2 = ?$$

Zmienna y jest tradycyjnie nazywana zmienną *objaśnianą* lub *zależną*. Zmienne x nazywa się zmiennymi *objaśniającymi* lub *niezależnymi*. Zarówno zmienne objaśniane i objaśniające mogą być wielkościami skalarnymi lub wektorami. Niech $(X_i, Y_i), i = 1, 2, \ldots, n$ — wartości doświadczalne ($x_i \neq x_j$ dla $i \neq j$). Szukamy $y = \hat{\beta}_1 + \hat{\beta}_2 x = f(x)$, która by przechodziła możliwie najbliżej wszystkich punktów doświadczalnych (X_i, Y_i).

Metoda najmniejszych kwadratów 1

W przypadku metody najmniejszych kwadratów dopasowanie polega na minimalizacji sumy:

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} [Y_i - f(X_i)]^2 = \sum_{i=1}^{n} [Y_i - (\beta_1 + \beta_2 X_i)]^2.$$

Wiadomo, że funkcja wielu zmiennych ma minimum w punkcie, dla którego pochodne cząstkowe tej funkcji po wszystkich zmiennych są równe zeru, a zatem w tym przypadku muszą być spełnione warunki

$$\begin{cases} \dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} = 0 \\ \dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} = 0 \end{cases}$$

Metoda najmniejszych kwadratów 2

$$\begin{cases} \dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} = 0 \\ \dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_{i=1}^{n}(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)) = 0 \\ -2\sum_{i=1}^{n}(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i))X_i = 0 \end{cases} \Leftrightarrow$$

$$\begin{cases} \hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases}$$

Metoda najmniejszych kwadratów 3

Skąd według Wzory Cramera (dla rozwiązania układa równań przy pomocy wyznaczników) dla β_2 wynika:

$$\begin{cases} \hat{\beta}_2 = \dfrac{{}_n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{{}_n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \\ \hat{\beta}_1 = \dfrac{1}{n}\left(\sum_{i=1}^{n} Y_i - \hat{\beta}_2\left(\sum_{i=1}^{n} X_i\right)\right) \end{cases}$$

Oszacowana wartość błędu w takim przypadku wynosi:

$$\hat{\sigma} = \frac{1}{n}\left(\sum_{i=1}^{n}(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i))^2\right)$$

Twierdzenie Gaussa-Markowa

Jeżeli próba posiada następujące właściwości model danych jest poprawnie określony, wtedy estymator najmniejszych kwadratów jest najlepszym (tj. mającym najmniejszą wariancję) estymatorem spośród liniowych, nieobciążonych estymatorów liniowego modelu regresji (ang. *best linear unbiased estimator, BLUE*).

Przykład

Dane są w tablice. Obliczyć regresją liniową.

Pojemność RAM	Cena
2	12
4	16
8	28
16	62

Tablica: Dane RAM

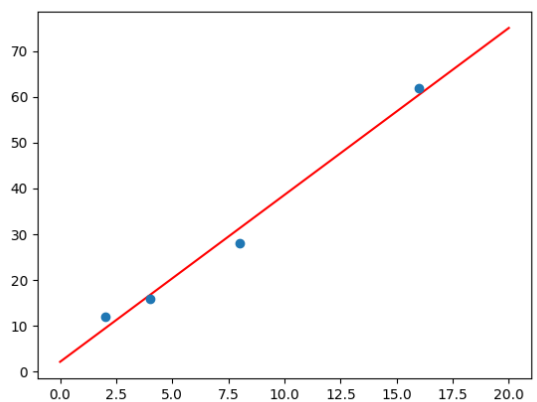
Rozwiązanie

$$\hat{\beta}_2 = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{4(24+64+224+992) - 30 \cdot 118}{4(340) - 900}$$

$$= \frac{1676}{460} \approx 3.643$$

oraz

$$\hat{\beta}_1 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - \hat{\beta}_2 \left(\sum_{i=1}^{n} X_i\right)\right) = \frac{1}{4}\left(118 - \frac{1676}{460} \cdot 30\right) \approx 2.17$$



Python

```
import numpy as np
import matplotlib.pyplot as plt

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])
n = len(RAM)
beta2 = (n*sum(RAM*price) - sum(RAM)*sum(price))/
              (n*sum(RAM**2) - (sum(RAM))**2)
print(beta2)
beta1 = (sum(price)-beta2*sum(RAM))/n
print(beta1)

x = np.linspace(0,20)
y = beta1 + x*beta2
plt.plot(x,y,'r')
plt.plot(RAM, price,'o')
plt.show()
```

Python: stats

```
import numpy as np
from scipy import stats

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])

print(np.polyfit(RAM,price,1))
#lub zaawansowanej
reg= stats.linregress(RAM, price)
print(reg)
```

Obrazek taki sam

Współczynnik korelacje

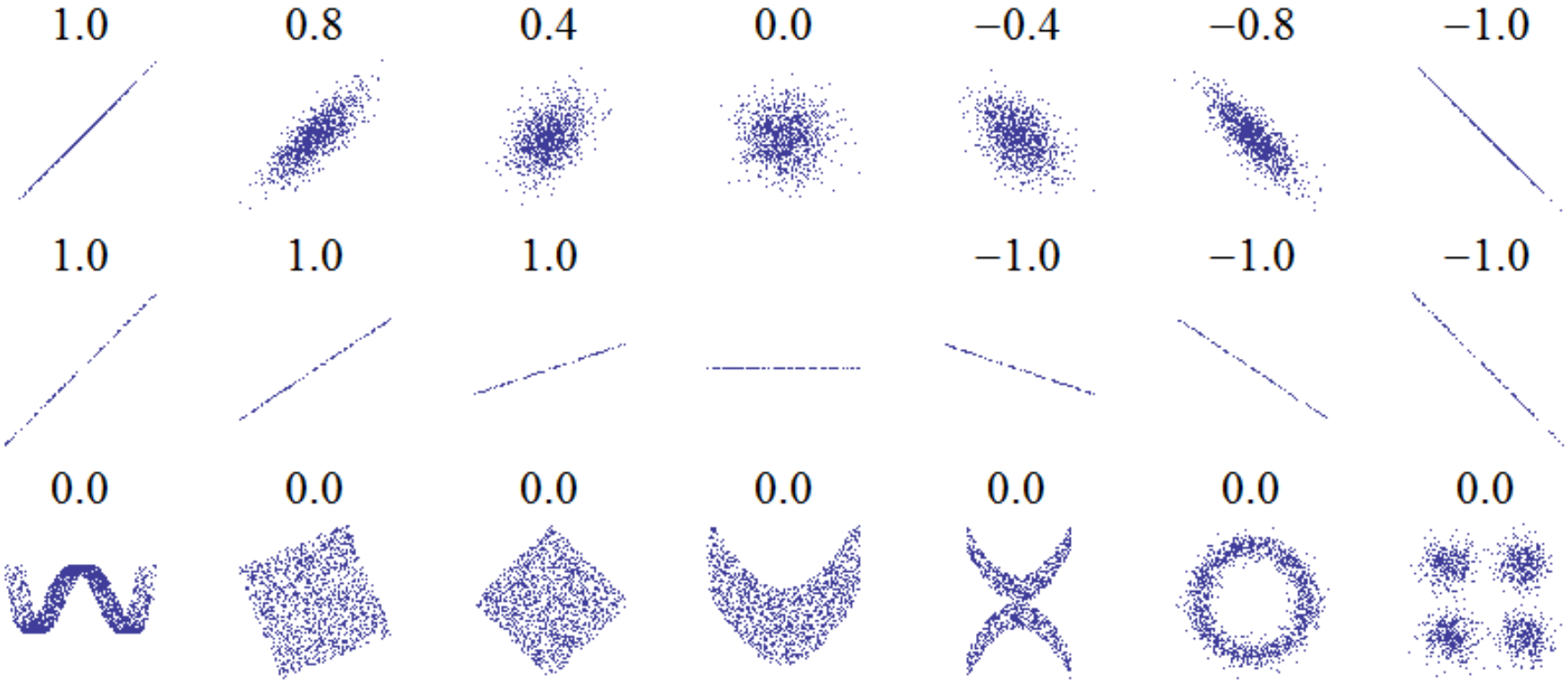
Współczynnik korelacji

– liczba określająca, w jakim stopniu zmienne są współzależne. Jest to miara korelacji dwóch (lub więcej) zmiennych. Istnieje wiele różnych wzorów określanych jako współczynniki korelacji. Większość z nich jest normalizowana tak, żeby przybierała wartości od -1 (zupełna korelacja ujemna), przez 0 (brak korelacji) do $+1$ (zupełna korelacja dodatnia).

Najczęściej stosowany jest *współczynnik korelacji r Pearsona*. W przypadku rozkładu dalekiego od dwuwymiarowego normalnego lub istnienia w próbie obserwacji odstających współczynnik korelacji Pearsona może fałszywie wskazywać na nieistniejącą korelację (zjawisko to widać na przykładzie kwartetu Anscombe'a).

Współczynnik korelacji liniowej Pearsona

Współczynnik korelacji liniowej Pearsona (r) – współczynnik określający poziom zależności liniowej między zmiennymi losowymi. Został opracowany przez Karla Pearsona.



Źródło: Autorstwa Imagecreator z angielskiej Wikipedii - Ten diagram został stworzony za pomocą Mathematica, Domena publiczna, `https://commons.wikimedia.org/w/index.php?curid=3732359`

Współczynnik korelacji liniowej Pearsona: wzór

Niech x i y będą zmiennymi losowymi o dyskretnych rozkładach. x_i i y_i oznaczają wartości prób losowych tych zmiennych ($i = 1, 2, \ldots, n$), natomiast $\overline{x}$, $\overline{y}$ – wartości średnie z tych prób, tj.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Wówczas współczynnik korelacji liniowej definiuje się następująco:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

gdzie $r_{xy} \in [-1, 1]$.
Ogólnie współczynnik korelacji liniowej dwóch zmiennych jest ilorazem kowariancji i iloczynu odchyleń standardowych tych zmiennych:

$$r_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Przykład

Dane są w tablice. Obliczyć współczynnik korelacji liniowej

Pojemność RAM	Cena
2	12
4	16
8	28
16	62

Tablica: Dane RAM

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 7.5, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 29.5.$$

Wówczas współczynnik korelacji liniowej definiuje się następująco:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

$$= \frac{(-5.5)(-17.5) + (-3.5)(-13.5) + (0.5)(-1.5) + (8.5)(32.5)}{\sqrt{(-5.5)^2 + (-3.5)^2 + (0.5)^2 + (8.5)^2}\sqrt{(-17.5)^2 + (-13.5)^2 + (-1.5)^2 + (8)^2}}$$

$$= \frac{419}{\sqrt{115 * 1547.0}} = 0.99,$$

Python: numpy r

```
import numpy as np

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])


mRAM = np.mean(RAM)
mprice = np.mean(price)
nRAM = RAM - np.mean(RAM)
nprice = price - np.mean(price)
r = sum(nRAM*nprice/(sum(nRAM**2)*sum(nprice**2))**0.5)
print(r)
```

Python: stats r

```
import numpy as np
from scipy import stats

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])

reg= stats.linregress(RAM, price)
print(reg.rvalue)
```

Poziomy korelacji i ich interpretacja

Korelacje można interpretować jako silne, słabe, ujemne. Interpretacja taka jest jednak arbitralna i nie możemy jej traktować zbyt ściśle. Na przykład współczynnik równy 0,9 dla socjologów i ekonomistów oznacza silną korelację, a dla fizyków posługujących się wysokiej klasy pomiarami przy badaniu praw przyrody oznacza korelację słabą.

Korelacje	**Ujemne**	**Dodatnie**
Słabe	$-0{,}5$ do $0{,}0$	$0{,}0$ do $0{,}5$
Silne	$-1{,}0$ do $-0{,}5$	$0{,}5$ do $1{,}0$

Uwaga! Matematyczne skorelowane mogą być zjawiska, który nie maja nic wspólnego pomiędzy sobą, zobacz stroną `https://www.tylervigen.com/spurious-correlations`.

Współczynnik determinacji oraz współczynnik zbieżności

Współczynnik determinacji R^2 jest zdefiniowane jako

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \geq 0,$$

gdzie:

- Y_i – i-ta obserwacja zmiennej Y,
- $\hat{Y}_i$ – wartość teoretyczna zmiennej objaśnianej (na podstawie modelu),
- $\overline{y}$ – średnia arytmetyczna empirycznych wartości zmiennej objaśnianej.

Współczynnik zbieżności definiuje się jako

$$\phi^2 = 1 - R^2.$$

Współczynnik determinacji oraz współczynnik korelacje Pearsona

$$R^2 = r_{xy}^2$$

Dowód: na tablice

Przykład

Dane są w tablice. Obliczyć współczynnik determinacji

Pojemność RAM	Cena
2	12
4	16
8	28
16	62

Tablica: Dane RAM

$$y = 2.17 + 3.643x$$

$$Y = [9.456, 16.742, 31.314, 60.458]$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 7.5, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 29.5.$$

Wówczas współczynnik determinacji definiuje się następująco:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{20^2 + 12.8^2 + 1.8^2 + 30.9^2}{(-17.5)^2 + (-13.5)^2 + (-1.5)^2 + (8)^2} = \frac{1521.88}{\sqrt{1547.0}} = 0.98.$$

Python: numpy R2

```
import numpy as np

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])
mprice = np.mean(price)
n = len(RAM)
beta2 = (n*sum(RAM*price) - sum(RAM)*sum(price))/
            (n*sum(RAM**2) - (sum(RAM))**2)
print(beta2)
beta1 = (sum(price)-beta2*sum(RAM))/n
print(beta1)

pricehat = beta1 + RAM*beta2
R2 = sum((pricehat-mprice)**2)/sum((price-mprice)**2)
print(R2)
#0.9868244287681628
nRAM = RAM - np.mean(RAM)
nprice = price - np.mean(price)
r = sum(nRAM*nprice/(sum(nRAM**2)*sum(nprice**2))**0.5)
print(r**2)
#0.9868244287681627
```

Python: stats R2

```
import numpy as np
from scipy import stats

RAM = np.array([2, 4, 8, 16])
price = np.array([12, 16, 28, 62])

reg= stats.linregress(RAM, price)
print(reg.rvalue**2)
#0.9868244287681629
```

Kwartet Anscombe'a

Kwartet Anscombe'a to cztery zestawy danych o identycznych cechach statystycznych, takich jak średnia arytmetyczna, wariancja, współczynnik korelacji czy równanie regresji liniowej, jednocześnie wyglądających zgoła różnie przy przedstawieniu graficznym. Układ tych danych został stworzony w 1973 roku przez brytyjskiego statystyka Francisa Anscombe'a aby ukazać znaczenie graficznej reprezentacji danych przy okazji ich analizy statystycznej.

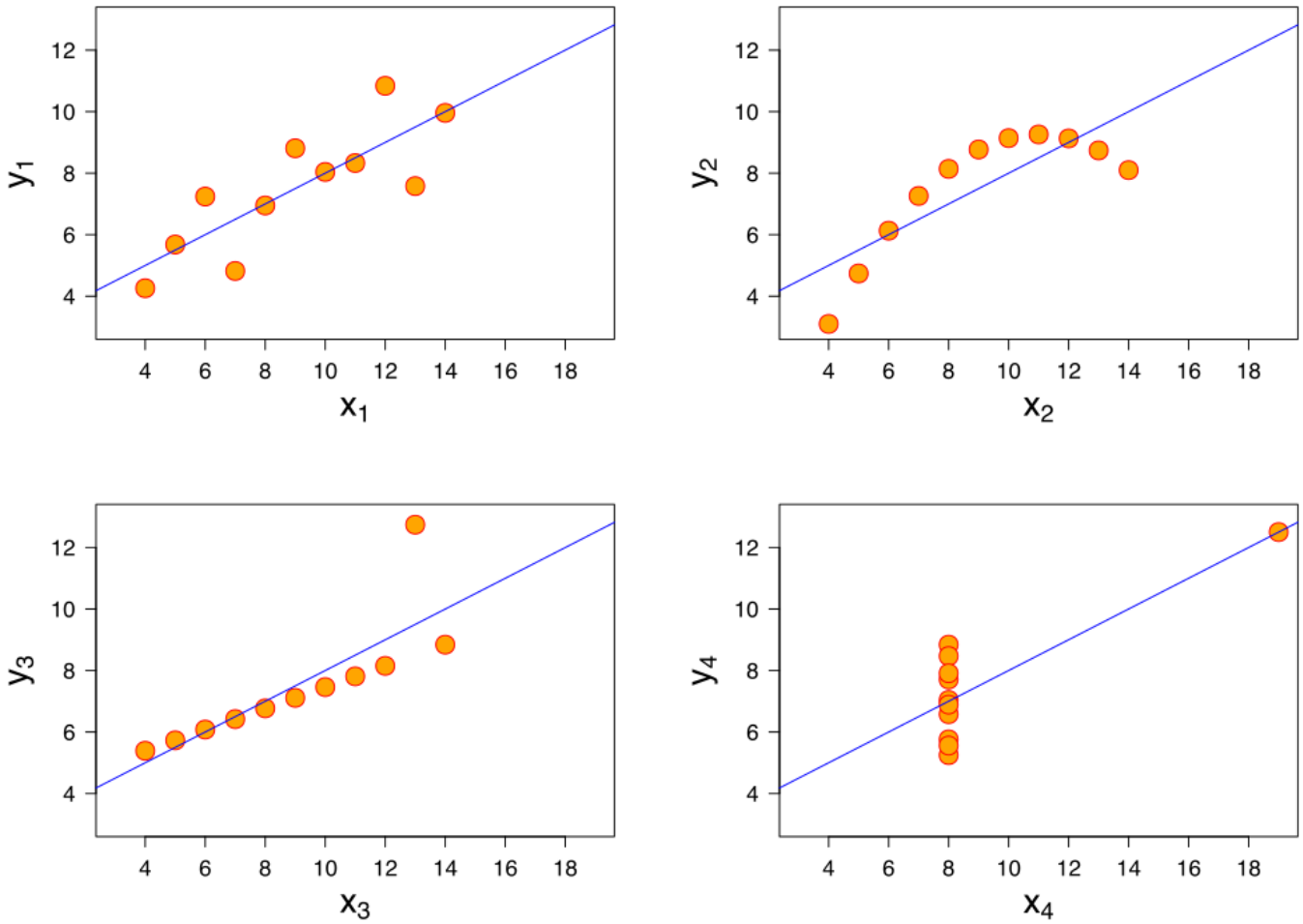
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Tablica: Kwartet Anscombe'a

Cechy układów

Cecha	Wartość	Dokładność
Średnia arytmetyczna zmiennej x	9	dokładnie
Wariancja zmiennej x	11	dokładnie
Średnia arytmetyczna zmiennej y	7.50	do 2 cyfr po przecinku
Wariancja zmiennej y	4.125	± 0.003
Współczynnik korelacji pomiędzy zmiennymi	0.816	do 3 cyfr po przecinku
Równanie regresji liniowej	$y = 3.00 + 0.500x$	do 2 i do 3 cyfr po przecinku
Współczynnik determinacji* R^2	0.67	do 2 cyfr po przecinku

Kwartet Anscombe'a



Rysunek: źródło: https://pl.wikipedia.org/wiki/Kwartet_Anscombe%E2%80%99a