REGRESJA LINIOWA

Pod pojęciem regresji należy rozumieć zależność zmiennej losowej Y od zmiennej losowej (lub deterministycznej) X postaci $Y=f(X)+e$, gdzie f jest pewną funkcją, zwaną funkcją regresji, e jest pewną zmienną losową o wartości oczekiwanej równej zeru, zaś X opisuje zespół parametrów charakteryzujących w pewien sposób warunki zewnętrzne.

Mówiąc prościej łatwo wyobrazić sobie sytuacje, gdy mierzymy dwie wielkości x i y związane ze sobą pewną zależnością. Wiemy jaki sens fizyczny ma taka zależność, a zatem łatwo określić czy jest ona liniowa, wielomianowa, logarytmiczna, wykładnicza czy też sinusoidalna. Problemem są nieznane na wstępie współczynniki występujące w ogólnych równaniach funkcji, a cała procedura sprowadza się do ich wyznaczenia na podstawie analizowanych danych. Służą do tego pewne funkcje w jakie wyposażone są arkusze kalkulacyjne.

Rozważamy sytuację, gdy mierzone dwie wielkości x i y związane są ze sobą równaniem liniowym:

$y = ax + b$

Wykonując pomiary uzyskujemy pary liczb (X,Y). Mając zbiór danych pomiarowych należy wyznaczyć stałe a i b tak, by znaleźć równanie linii prostej, najlepiej opisujące zależność pomiędzy X i Y. Dokonując tego zgodnie z **metodą najmniejszych kwadratów** oczekujemy, że sumaryczna odległość pomiędzy punktami pomierzonymi (X,Y) a punktami wyznaczonymi (x,y) ze znalezionego równania prostej będzie jak najmniejsza:

$$\sum_{i=1}^{n} \sqrt{(Y_i - y_i)^2 + (X_i - x_i)^2} = min$$

gdzie: n – liczba rozważanych punktów

Uwzględniając to, że odcięte kolejnych punktów są takie same czyli $x_i=X_i$, oraz wykorzystując równanie prostej wyrażenie upraszcza się do:

$$\sum_{i=1}^{n} \sqrt{(Y_i - (a \cdot X_i + b))^2} = min$$

gdzie a i b są empirycznymi współczynnikami regresji liniowej.

Równanie to przyjmuje wartość minimalną gdy współczynniki a i b obliczane są ze wzorów:

$$a = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - a\sum_{i=1}^{n} X_i\right)$$

Powyższe równania realizowane są w arkuszu kalkulacyjnym odpowiednio przez funkcje: *NACHYLENIE()* i *ODCIĘTA()*.

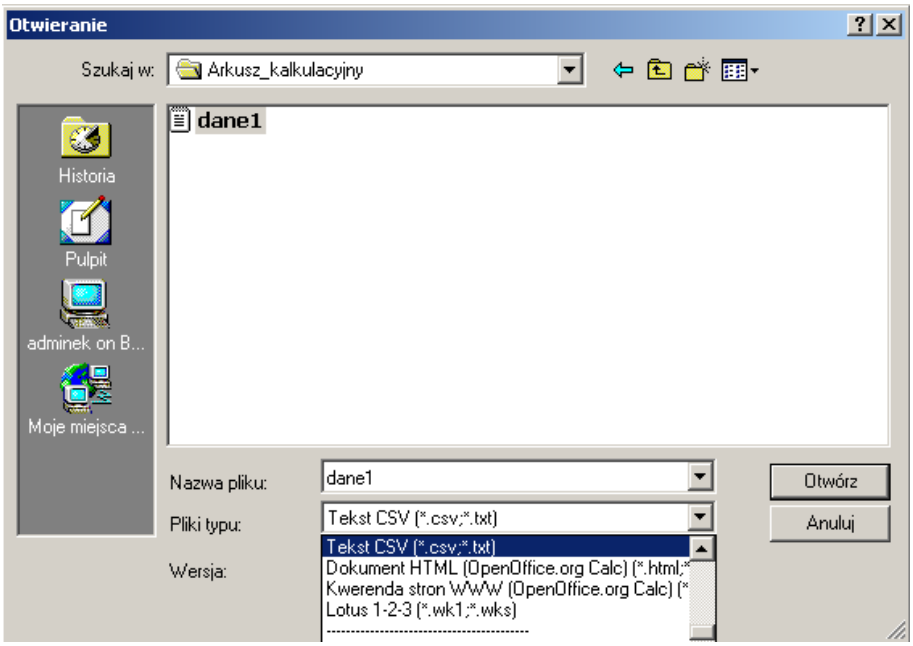
ĆWICZENIE

IMPORT DANYCH

Wczytywanie danych z pliku tekstowego

W arkuszu kalkulacyjnym OOo Calc:

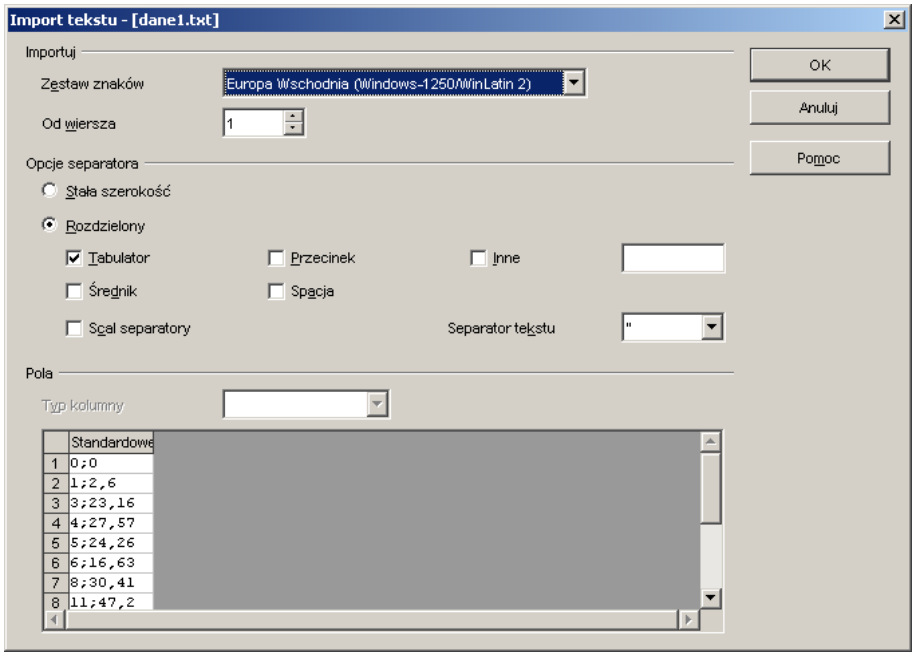
- menu *Plik* > *Otwórz*
- Wybrać odpowiedni katalog
- Wybrać odpowiedni typ pliku **(wg rysunku)**
- Otworzyć plik



Pliki tekstowe ***CSV*** (ang. Comma Separated Values) to pliki, w których wartości rozdzielone są przecinkiem. Jest to jeden z formatów przechowywania danych pochodzących z pomiarów laboratoryjnych. Według standardu wartości powinny być rozdzielone przecinkami, lecz dopuszcza się stosowanie innych znaków jak średniki czy tabulatory.

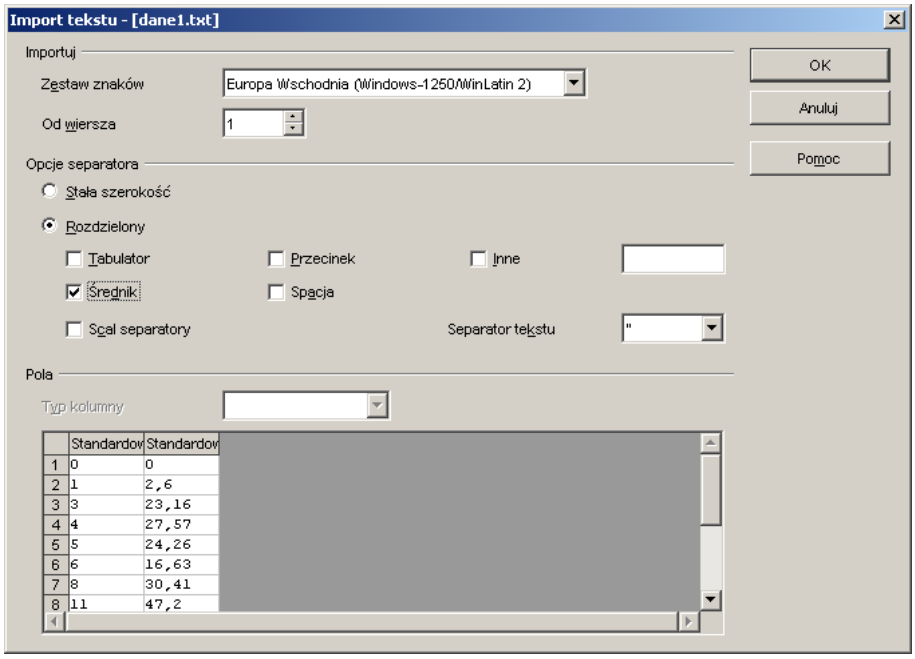
Pliki typu ***CSV*** mogą posiadać rozszerzenie ***txt***.

Ustawienia importu tekstu



W realizowanym przykładzie dane w pliku są rozdzielone za pomocą średników co widać na rysunku przedstawiającym okienko Importu tekstu. W takim wypadku należy włączyć odpowiedni rodzaj separatora: **średnik**.

Spowoduje to rozdzielenie na dwie kolumny danych czytanych z pliku tak, jak to widać na rysunku:



Na tym kończą się ustawienia dotyczące importu. Kliknięcie w **OK** zamyka okno "Import tekstu" a w arkuszu pojawiają się w kolumnach A i B dwa zbiory danych.

Separator dziesiętny (część wyłącznie informacyjna)

Kolejną rzeczą, na którą należy zwrócić uwagę podczas importowania danych tekstowych do arkusza kalkulacyjnego jest to jakim znakiem oddzielona jest część całkowita od części ułamkowej liczby. W zależności od programu rolę znaku rozdzielającego może pełnić **kropka** lub **przecinek**. Arkusz OOo Calc wykorzystuje w tym celu **przecinek**.

Jeśli w importowanych danych znakiem dziesiętnym byłaby **kropka**, to po zaimportowaniu zobaczymy w arkuszu:

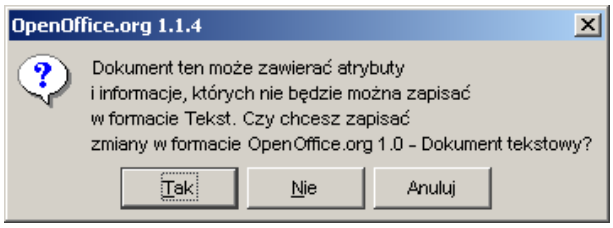
	A	B
1	0	0
2	1	02.06.05
3	3	23.16
4	4	27.57
5	5	24.26
6	6	16.63
7	8	30.41
8	11	47.2
9	12	50.03
10	13	60.33
11	14	59.89
12	16	71.18
13	17	84.27
14	19	77.69

Jak widać druga z importowanych wartości w drugiej kolumnie została rozpoznana jako data. Pozostałe dane z kolumny B (oprócz zero) są traktowane jako tekst (łatwo to stwierdzić na podstawie wyrównania do lewej krawędzi komórki) i jako takie nie mogą być wykorzystywane do obliczeń.

Zmiany znaku dziesiętnego można dokonać chociażby w edytorze tekstu OpenOffice.org Writer. W tym celu należy:

- otworzyć plik z danymi: menu *Plik* > *Otwórz* > *Pliki Typu*: "Dane tekstowe"
- menu *Edycja* > *Znajdź i Zamień*
- w polu *Szukaj* wpisać: **.**
- w polu *Zamień na* wpisać: **,**
- kliknąć w przycisk *Zamień wszyst*.
- zapisać plik z danymi: menu *Plik* > *Zapisz*.

Podczas operacji zapisywania zobaczymy komunikat:

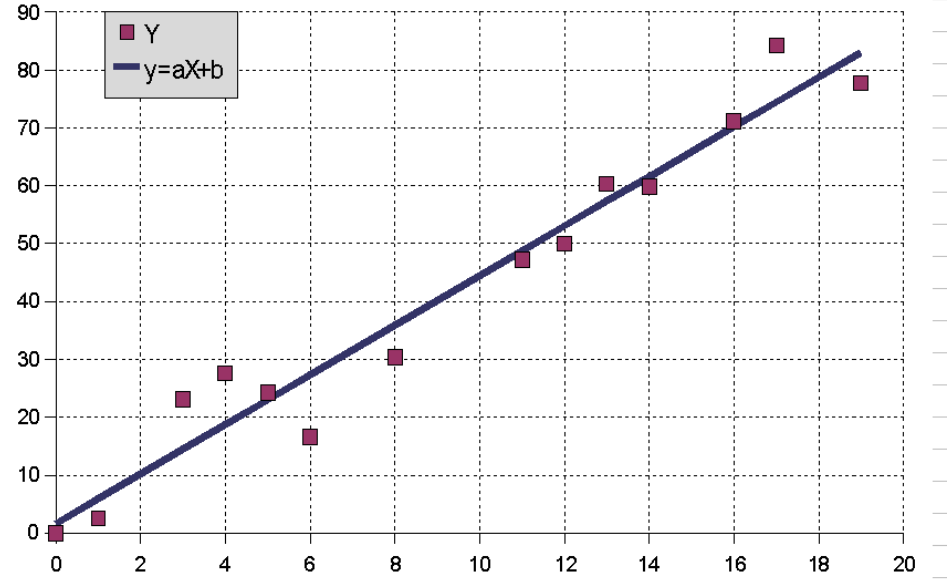


W celu zachowania danych jako pliku tekstowego **.txt* (który można importować do arkusza kalkulacyjnego) należy wybrać NIE. Opcja TAK oznacza zapisanie danych w pliku **.sxw*.

OBLICZENIA

Wyznaczyć współczynniki ***a*** i ***b*** regresji liniowej zaimportowanych danych oraz sporządzić wykres jak na wzorcu:

	A	B	C
1	***X***	***Y***	***y=aX+b***
2	0	0	1,59
3	1	2,6	5,87
4	3	23,16	14,45
5	4	27,57	18,73
6	5	24,26	23,02
7	6	16,63	27,31
8	8	30,41	35,88
9	11	47,2	48,74
10	12	50,03	53,03
11	13	60,33	57,32
12	14	59,89	61,6
13	16	71,18	70,18
14	17	84,27	74,46
15	19	77,69	83,04
16			
17		***a***=	4,287
18		***b***=	1,587
19			



1. Obliczyć wartość współczynnika ***a*** – funkcja *NACHYLENIE()* oraz wartość współczynnika ***b*** – funkcja *ODCIĘTA()*
2. Obliczyć wartości ***y*** z równania $\boldsymbol{y = aX + b}$, na podstawie *X* z wykorzystaniem wyznaczonych już stałych ***a*** i ***b***
3. Sporządzić wykres typu **Punktowy *(XY)***, na którym za pomocą punktów o kwadratowym znaczniku będą prezentowane dane pomiarowe czyli dane z **kolumny B**. Niebieską linią przedstawić drugą serię danych czyli linię regresji liniowej jako ilustrację danych znajdujących się w **kolumnie C**.

Poniżej prezentowany jest efekt użycia wykresu typu ***Liniowy***. Wykres zbudowany jest w oparciu o te same dane – jego nieprzydatność do tego typu zadań jest jasno widoczna i zatem oczywista!