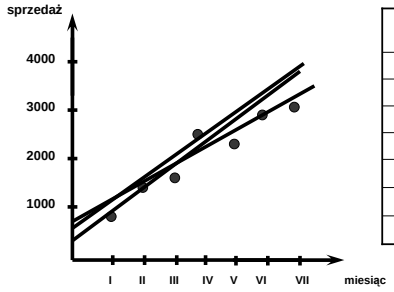
Hurtownie danych - przegląd technologii

Robert Wrembel
Politechnika Poznańska
Instytut Informatyki
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

Przewidywanie trendów

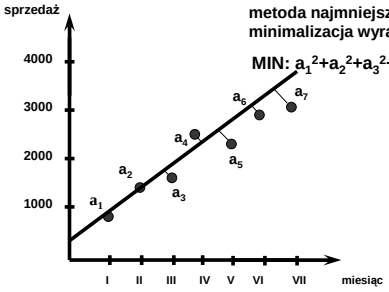
- Prosty mechanizm przewidywania bazujący na regresji
- Dostępny za pomocą zbioru funkcji języka SQL

Problematyka (1)



miesiąc	sprzedaż
I	1100
II	1850
III	2050
IV	3010
V	2860
VI	3200
VII	3840

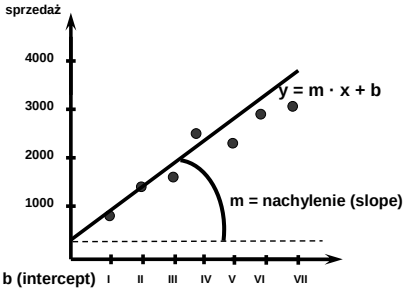
Problematyka (2)



metoda najmniejszych kwadratów: minimalizacja wyrażenia

MIN: $a_1^2+a_2^2+a_3^2+a_4^2+a_5^2+a_6^2+a_7^2$

Problematyka (3)



Problematyka (4)

- Metoda najmniejszych kwadratów
- Równanie prostej

$$y = m \cdot x + b$$

- Obliczenie współczynników m i b

$$m = \frac{n \cdot \sum_i (x_i \cdot y_i) - \sum_i x_i \cdot \sum_i y_i}{n \cdot \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$b = \frac{\sum_i y_i - m \cdot \sum_i x_i}{n}$$

Funkcje SQL

- Obliczanie współczynników m i b
- Nachylenie (m): REGR_SLOPE
- Przecięcie z Y (b): REGR_INTERCEPT

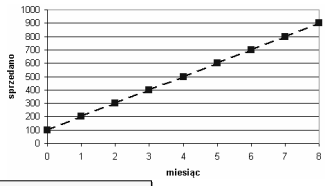
```
REGR_SLOPE (wyrażenie1 określające Y, wyrażenie2 określające X)
```

```
REGR_INTERCEPT (wyrażenie1, wyrażenie2)
```

- wyrażenie1 – wartości zmiennej zależnej (y)
- wyrażenie2 – wartości zmiennej niezależnej (x)
- Wartości te powinny być typu numerycznego lub transformowalnego do numerycznego

Prosty przykład (1)

MIESIAC	LICZBA
1	200
2	300
3	400
4	500
5	600
6	700
7	800
8	900




```
SELECT
REGR_SLOPE(liczba, miesiac) AS slope,
REGR_INTERCEPT (liczba, miesiac) AS intercept
FROM regr_data1;
```

SLOPE	INTERCEPT
100	100

Prosty przykład (2)

- Szacowana sprzedaż w 10-tym miesiącu

```
sprzedano(10) = REGR_SLOPE*10 + REGR_INTERCEPT
```

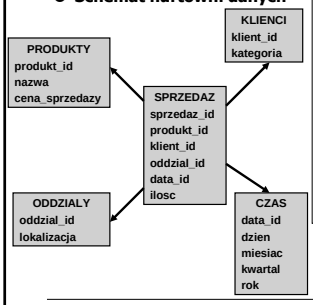
```
sprzedano(10) = 100*10 + 100 = 1100
```

```
select REGR_SLOPE(liczba, miesiac)*10 +
       REGR_INTERCEPT(liczba, miesiac) "sprzedano X"
from regr_data1;
```

sprzedano X
1100

Przykład złożony (1)

- Schemat hurtowni danych



l_sztuk	nazwa	cena	dzien	mies.	rok
40	Monitor	950	1	8	2004
21	Drukarka	300	13	8	2004
12	Monitor	950	16	8	2004
33	Napęd CD	70	21	8	2004
14	Napęd CD	70	29	8	2004
22	Napęd CD	70	2	9	2004
34	Drukarka	300	4	9	2004
35	Monitor	950	13	9	2004
42	Drukarka	300	17	9	2004
13	Napęd CD	70	22	9	2004
24	Drukarka	300	25	9	2004
13	Monitor	950	3	10	2004
21	Drukarka	300	5	10	2004
32	Monitor	950	14	10	2004
23	Napęd CD	70	23	10	2004
64	Drukarka	300	30	10	2004
35	Napęd CD	70	1	11	2004
65	Drukarka	300	2	11	2004

Przykład złożony (2)

- Wyznaczyć linię trendu sprzedaży produktów (kwoty sprzedaży) dla ostatnich czterech miesięcy roku 2004

```
SELECT
  REGR_SLOPE(SUM(l_sztuk*cena_sprzedazy), miesiac) AS slope,
  REGR_INTERCEPT (SUM(l_sztuk*cena_sprzedazy), miesiac) AS intercept
FROM sprzedaz NATURAL JOIN produkty NATURAL JOIN czas
WHERE rok=2004 AND miesiac BETWEEN 8 AND 11
GROUP BY miesiac;
```

SLOPE	INTERCEPT
114695	-930985

SPRZEDANO	ROK	MIESIAC
58990	2004	8
65700	2004	9
69860	2004	10
439920	2004	11

Przykład złożony (3)

- Oszacowanie sprzedaży w grudniu 2004

```
SELECT
  REGR_SLOPE(SUM(l_sztuk*cena_sprzedazy), miesiac) *12 +
  REGR_INTERCEPT (SUM(l_sztuk*cena_sprzedazy), miesiac) AS grudzien
FROM sprzedaz NATURAL JOIN produkty NATURAL JOIN czas
WHERE rok=2004 AND miesiac BETWEEN 8 AND 11
GROUP BY miesiac;
```

GRUDZIEN
445355

SPRZEDANO	ROK	MIESIAC
58990	2004	8
65700	2004	9
69860	2004	10
439920	2004	11
445355	2004	12

Przykład złożony (4)

- Wyznaczenie linii trendu sprzedaży produktu 20 w kolejnych miesiącach roku 2004

```
SELECT miesiac,
       ROUND(REGR_SLOPE(l_sztuk, dzien), 2)  AS SLOPE
FROM Sprzedaz NATURAL JOIN Czas NATURAL JOIN produkty
WHERE rok=2004 AND produkt_id=20
GROUP BY miesiac;
```

MIESIAC	SLOPE	
8		
9	-,37	↘
10	1,72	↗
11	-,14	↘

Przykład złożony (5)

- Wyznaczenie linii trendu sprzedaży produktu 20 w poszczególnych miesiącach roku 2004, w każdym z oddziałów firmy

```
SELECT miesiac, oddzial_id,
       ROUND(REGR_SLOPE(l_sztuk, dzien), 2) as slope
FROM Sprzedaz NATURAL JOIN Czas NATURAL JOIN produkty
WHERE rok=2004 AND produkt_id=20
GROUP BY miesiac, oddzial_id;
```

MIESIAC	ODDZIAL_ID	SLOPE
8	41	
9	40	
9	41	-2,25
10	40	-1,78
10	41	3,56
11	40	-1,76
11	41	-,94

Wielkość wykorzystanej próbki

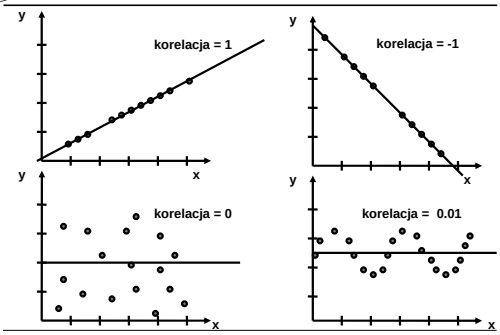
- Funkcje REGR_SLOPE i REGR_INTERCEPT eliminują dane, których pierwszy lub drugi argument przyjmuje wartość pustą
- Liczbę rekordów o obu wartościach niepustych wykorzystanych do wyliczenia współczynników wyznacza funkcja REGR_COUNT

```
REGR_COUNT (wyrażenie1, wyrażenie2)
```

```
SELECT miesiac,
REGR_COUNT(l_sztuk, dzien) AS REGR_COUNT
FROM Sprzedaz NATURAL JOIN Czas NATURAL JOIN produkty
WHERE rok=2004 AND produkt_id=20
GROUP BY miesiac;
```

MIESIAC	REGR_COUNT
8	1
9	3
10	4
11	9

Jakość regresji liniowej



Współczynnik korelacji (1)

- Jest miarą jakości regresji liniowej ⇨ miarą liniowej zależności dwóch zmiennych
- Współczynnik standardowy:
  - wartości z przedziału < -1 , 1 >
  - wartości 1 oraz -1 oznaczają idealną zgodność
  - wartość 0 oznacza brak zależności
- Współczynnik R2:
  - kwadrat współczynnika standardowego
  - wartości z przedziału < 0 , 1 >
  - wartość 1 oznacza idealną zgodność
  - wartość 0 oznacza brak zależności

Współczynnik korelacji (2)

```
SELECT
REGR_SLOPE(SUM(l_sztuk*cena_sprzedazy), miesiac) AS slope,
REGR_INTERCEPT (SUM(l_sztuk*cena_sprzedazy), miesiac) AS intercept,
REGR_R2 (SUM(l_sztuk*cena_sprzedazy), miesiac) AS r2
FROM sprzedaz NATURAL JOIN produkty NATURAL JOIN czas
WHERE rok=2004 AND miesiac BETWEEN 8 AND 11
GROUP BY miesiac;
```

SLOPE	INTERCEPT	R2
114695	-930985	,623053708