Metody predykcji → analiza regresji wielokrotnej, nieliniowej i wybór zmiennych

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

TPD – 2009/2010
Aktualizacja w 2010

Przebieg wykładu

Poprzednie wykłady

1. Predykcja z wykorzystaniem analizy regresji
2. Ocena poprawności modelu regresji liniowej
3. Regresja wielowymiarowa
4. Ocena modeli predykcyjnych
5. Regresja nieliniowa
6. Selekcja zmiennych.
7. Inne podejścia do predykcji

Przypomnijmy ….

- Spróbujmy wspólnie powtórzyć co wiemy na temat tworzenia i diagnostyki modelu regresji!
- Także dla regresji wielowymiarowej
- Patrz klasyczne środki prezentacji

Ogólny schemat postępowania

Ustalenie założeń i postaci modelu

↓

Szacowanie parametrów strukturalnych modelu na podstawie wyników próby

↓

Weryfikacja modelu:
- czy parametry są istotne?
- ocena założeń modelu (reszty)
- czy zależność jest liniowa?

NIE → Ustalenie założeń i postaci modelu

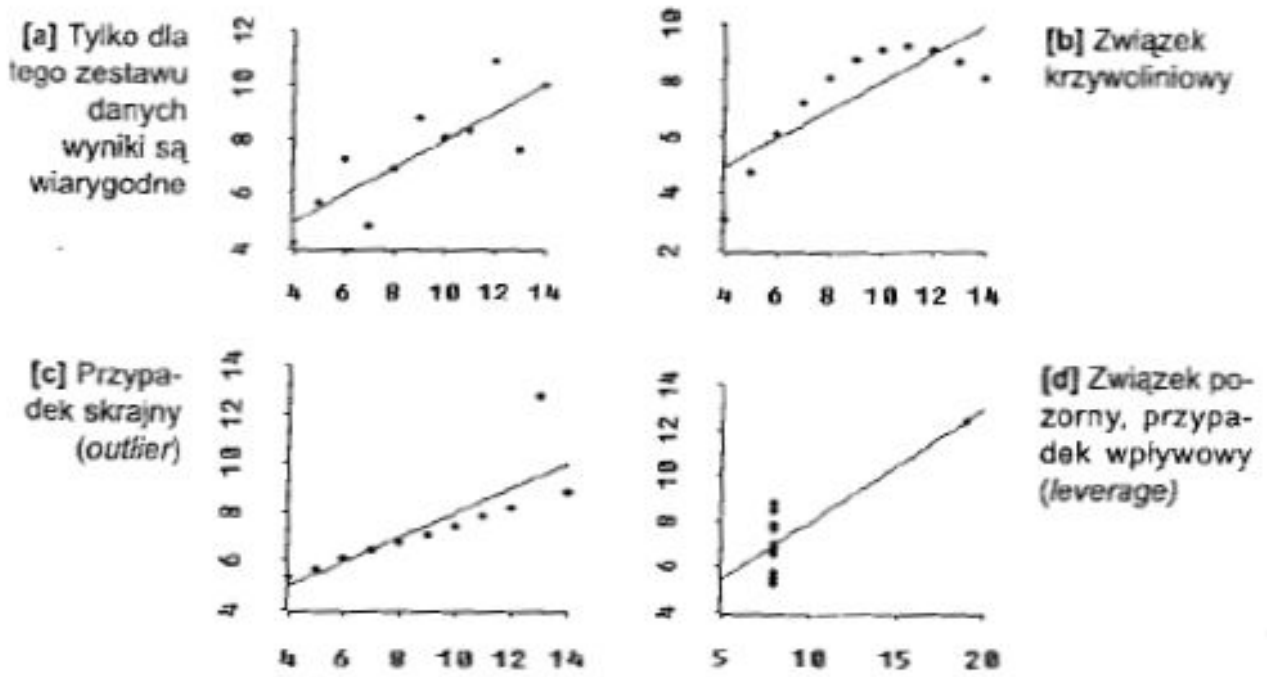
↓ **TAK**

Wykorzystanie modelu:
1. Predykcja zmiennej Y
2. Opis zależności między zmiennymi

Trudne przypadki w modelach regresji

8.4. Problemy w interpretacji współczynnika korelacji

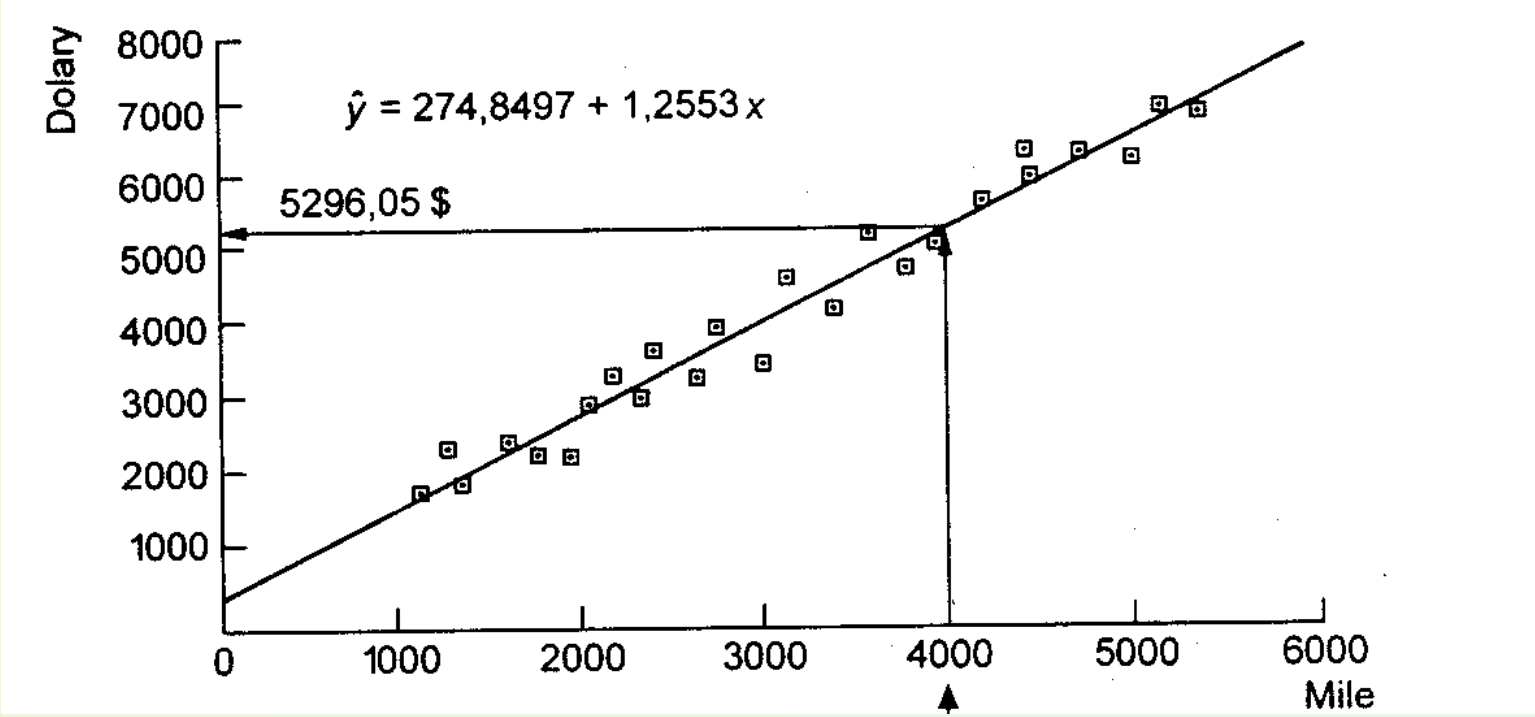
Na rysunku 8.6 przedstawiono wykresy korelacyjne czterech różnych grup wyników, dla których współczynnik korelacji wyników jest taki sam i wynosi $r = 0{,}816$. We wszystkich przypadkach zmienne mają takie same średnie $M_X = 9$ $M_Y = 7{,}5$, równanie regresji jest dokładnie takie samo. $Y' = 3 + 0{,}5 \times X$.



Rysunek 8.6. Przykład danych Anscombe'a.

Prognoza punktowa w regresji

- Łatwa na podstawie równania regresji.
- Np. oceń obciążenie kart wśród posiadaczy kart, których trasa podróży osiągnie 4000 mil, w okresie o takiej długości jak okres badany: $\hat{y} = 274{,}85 + 1{,}2663 \cdot x = 274{,}85 + 1{,}2663 \cdot 4000 = 5296{,}05$



Przedziały predykcji

- (1-α)·100% przedział predykcji zmiennej Y

$$\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

- Rozpiętość przedziału predykcji zależy od odległości wartości *x* od średniej $\bar{x}$!

Przykład: posiadacz, który przebył 4000 mil i 95% przedział ufności.

- Z analizy danych historycznych:

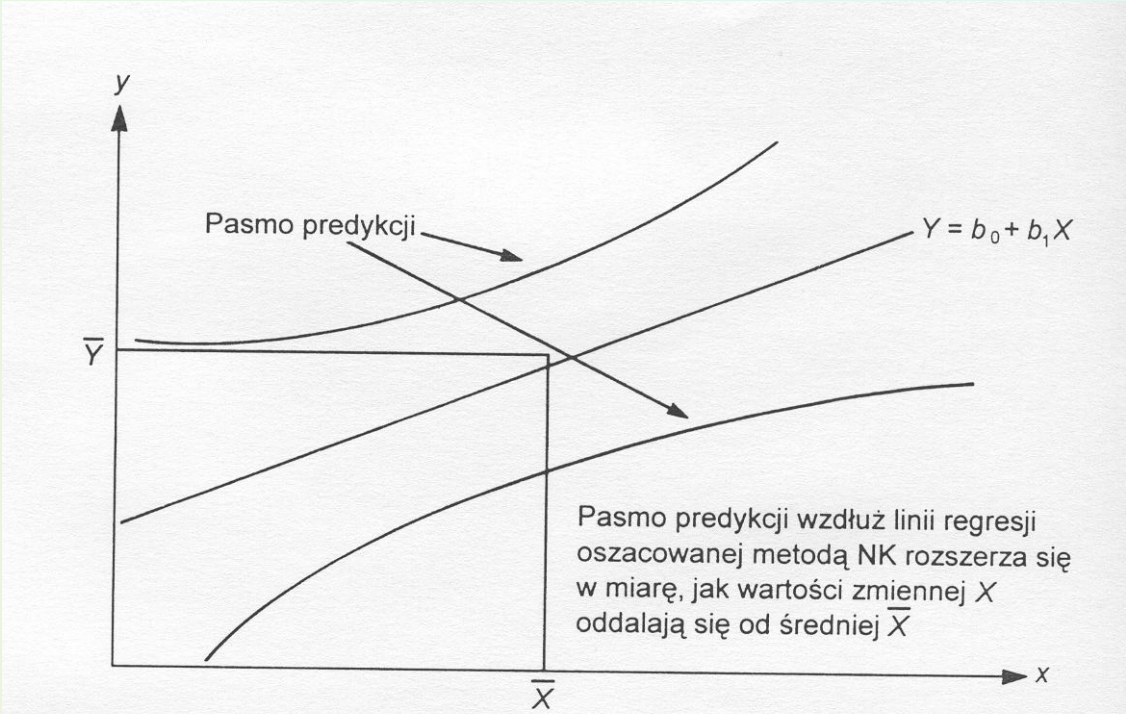
  $\bar{x}$ = 79448/25=3177,92; SS_x = 40947557,84 a *s* = 318,16
  Ponadto *t* przy 23 stopniach swobody wynosi 2,069

Stąd przedział 5296,05$\pm$676,62 = [4619,43; 5972,67]

- Oznacza to, że w oparciu o wyniki badań można mieć 95% zaufania do prognozy, że posiadacz karty, który przebył trasę 4000 mil w okresie o danej długości obciąży swoją kartę kredytową sumą od 4619.43 do 5972,67$.

Niepewność w predykcji

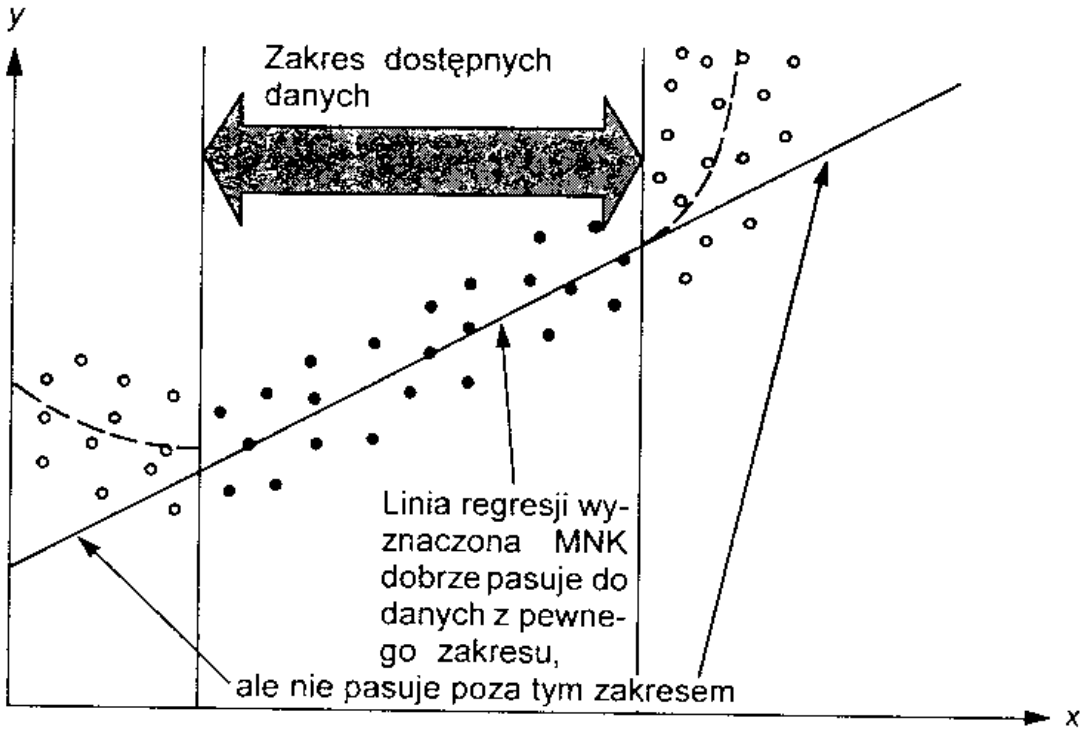
- Ograniczenie prognoz punktowych → błędu pochodzące zarówno z niepewności szacunków, jak i losowej zmienności położenia punktów w stosunku do linii regresji.
- Stosuj wtedy tzw. przedziały predykcji (tzw. prognozy przedziałowe).



Rysunek 10.29. Pasmo predykcji

Zakres przewidywania

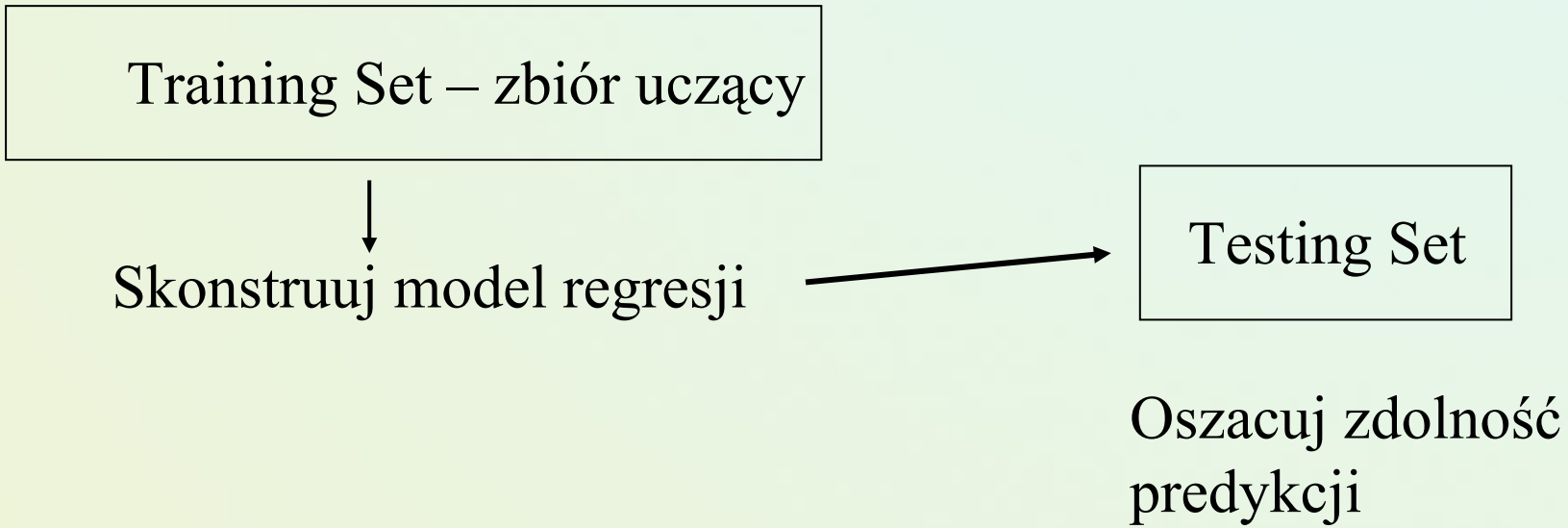
- Wartości prognozowane nie powinny zbyt wykraczać poza zakres wartości wykorzystywanych w procedurze szacowania parametrów równania regresji.



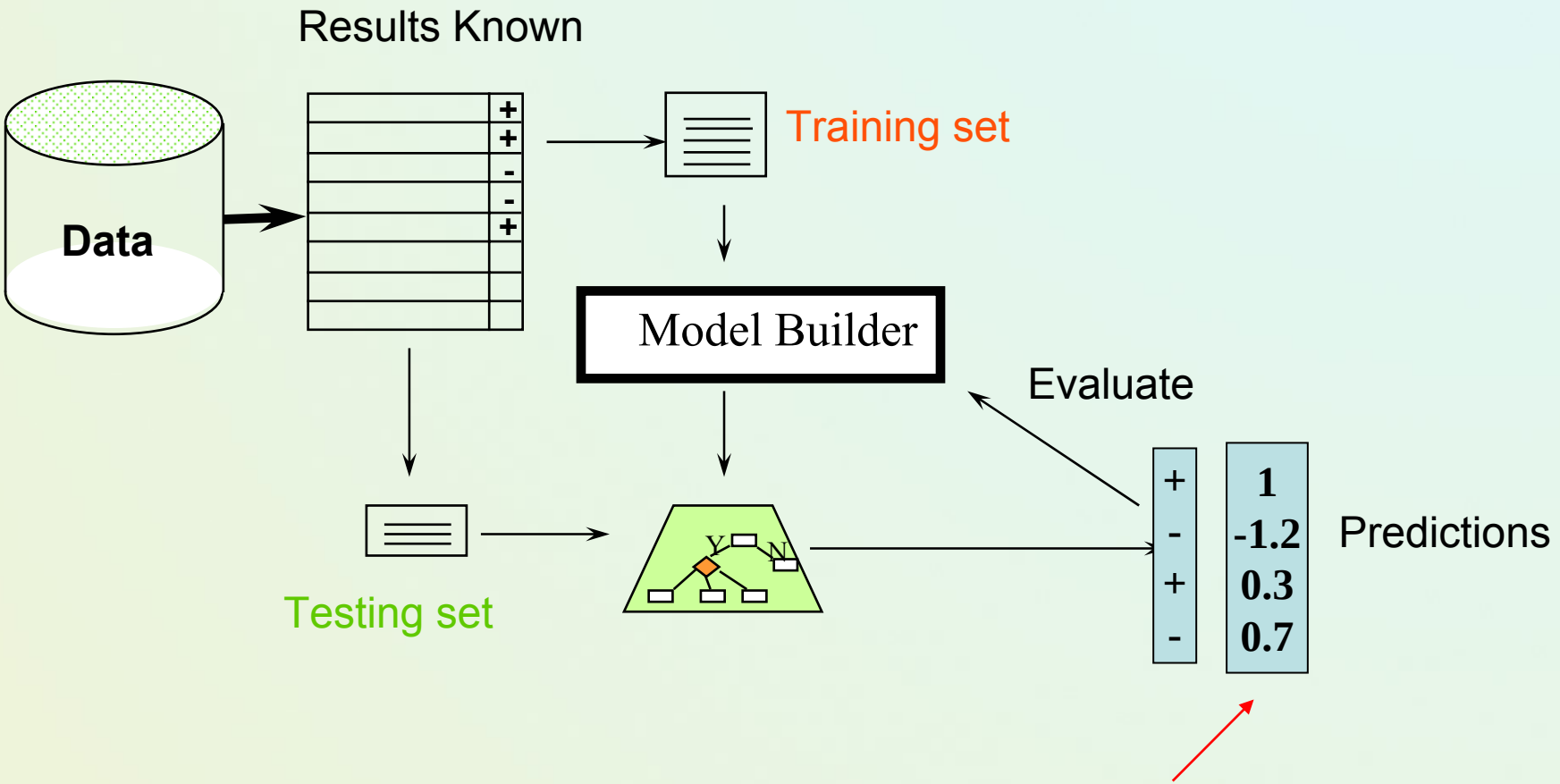
Rysunek 10.27. Niebezpieczeństwo ekstrapolacji

Inne spojrzenie na ocenę predykcji

- Przypomnijmy metody oceny klasyfikatorów
- Podejście „empiryczne”
- Podział danych na części ucząca vs. testowa



Podobne schematy podziału danych



W odróżnieniu od klasyfikacji nie ma etykiet dyskretnych klas, lecz wynik - wartość liczbową zmiennej zależnej

Ocena predykcji → „Predictor Error Measures”

- Pomiar wielkości różnicy między wartością rzeczywistą a przewidywaną
- **Loss function**: measures the error betw. y_i and the predicted value y_i^
  - Absolute error: $| y_i - y_i\hat{} |$
  - Squared error: $(y_i - y_i\hat{})^2$
- Test error (generalization error): the average loss over the test set (*n*)
  - Mean absolute error: $\frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$ Mean squared error: $\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$
  - Relative absolute error: $\frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{\sum_{i=1}^{n} | y_i - \bar{y} |}$ Relative squared error: $\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

The mean squared-error exaggerates the presence of outliers

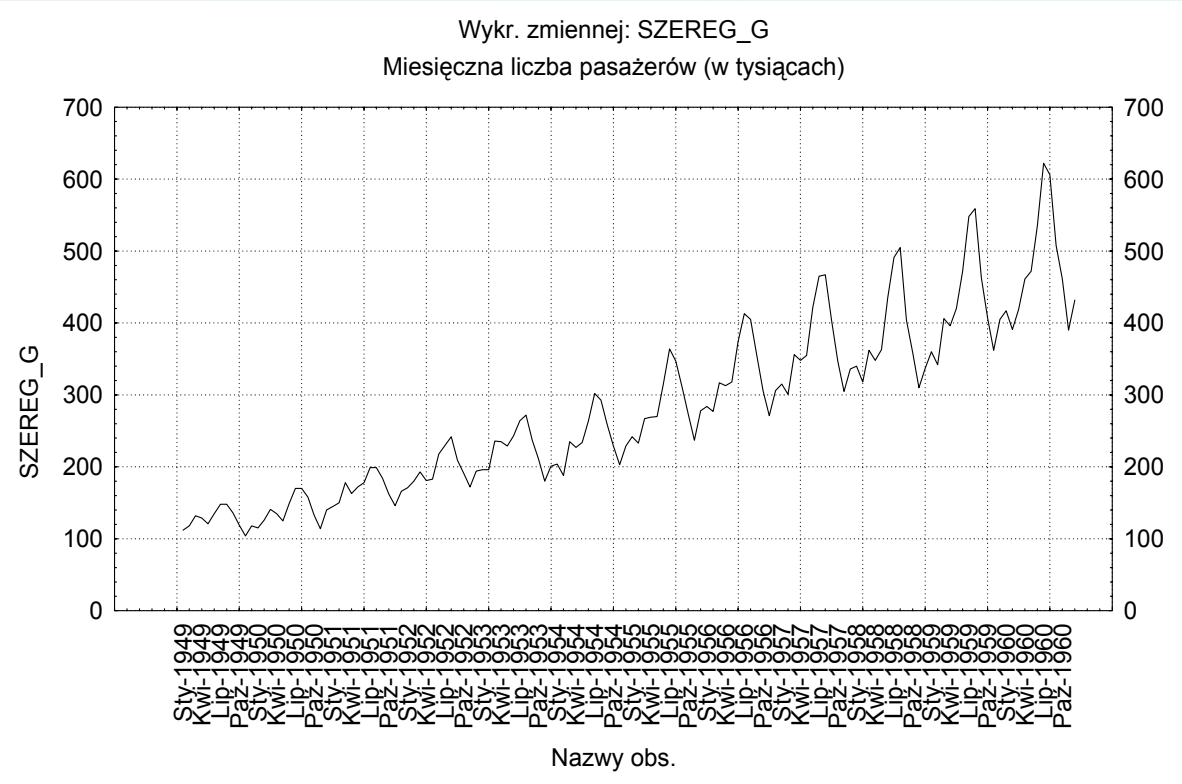
Popularly use (square) root mean-square error, similarly, root relative squared error

Predykcja w danych zmiennych czasowo – szeregi czasowe

Co to jest szereg czasowy?

- Ciąg obserwacji pewnego zjawiska w kolejnych jednostkach czasu.
- Tzw. funkcja trendu

- Predykcja
- Przesuwane okna



Inne przykłady regresji wielokrotnej

- Warto przeanalizować przykład 4.3. analizy samochodów podany przez Koronacki, Mielniczuk „Statystyka” rozdział 4.3.4
  - Przedstawiony opis obejmuje:
    - Tworzenie modelu regresji liniowej,
    - Diagnostykę modelu
    - Identyfikację obserwacji samotnicznych
    - Selekcje zmiennych

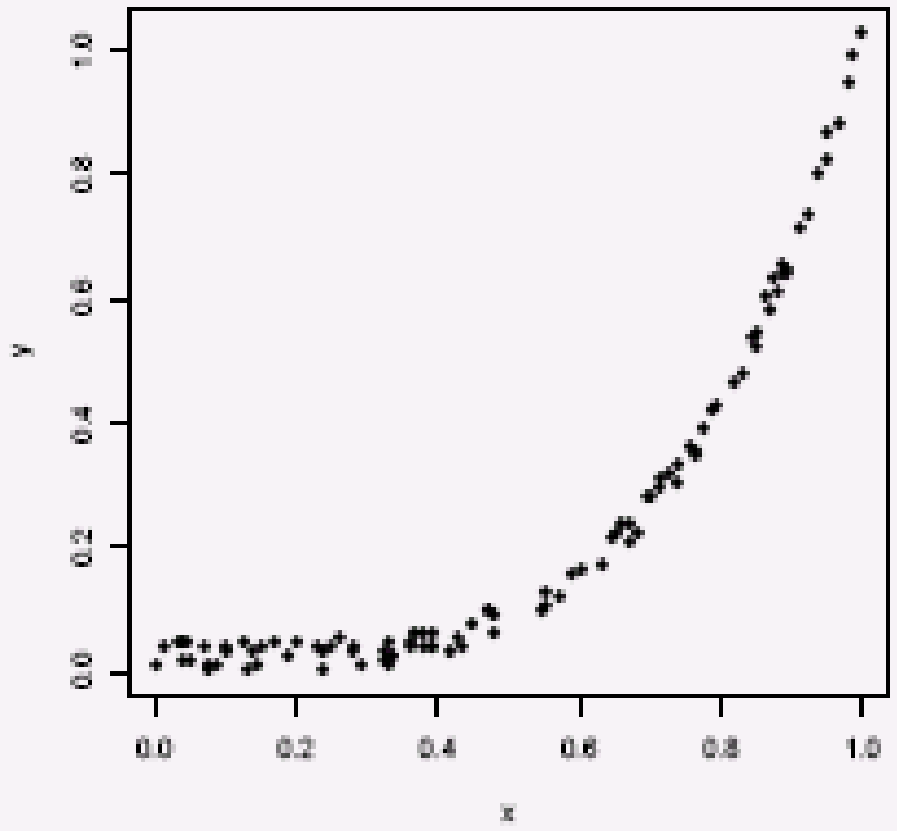
Założenia poprawności stosowania modelu regresji

- Zmienne niezależne **x** nie są ze sobą silnie skorelowane.
- Żadna ze zmiennych niezależnych nie powinna być kombinacją liniową innych zmiennych niezależnych.
- Liczba obserwacji *n* musi być większa od liczby parametrów do oszacowania
- Zakłada się istnienie modelu liniowego względem parametrów.

- Jeśli wiele z założeń jest niespełniony nie korzystaj z przedstawionych metod weryfikacji
    - Bardziej adekwatny skorygowany współczynnik determinacji (także stosowalny gdy nie ma wyrazu wolnego).

Regresja nieliniowa

Co zrobić gdy zależność pomiędzy zmiennymi wygląda na nieliniową?



Regresja nieliniowa i transformacje do modelu liniowego

- Między zmienną objaśnianą a zmiennymi objaśniającymi mogą zachodzić związki nieliniowe.
- W wielu przypadkach można dokonać transformacji do modelu liniowego poprzez odpowiednie przekształcenia zmiennych.
- Model $Y = f(X,b)$ jest liniowy względem parametrów, jeśli można go przedstawić jako *liniową* funkcję *jednoznacznych przekształceń X*, przy czym współczynniki tych przekształceń musza być znane.

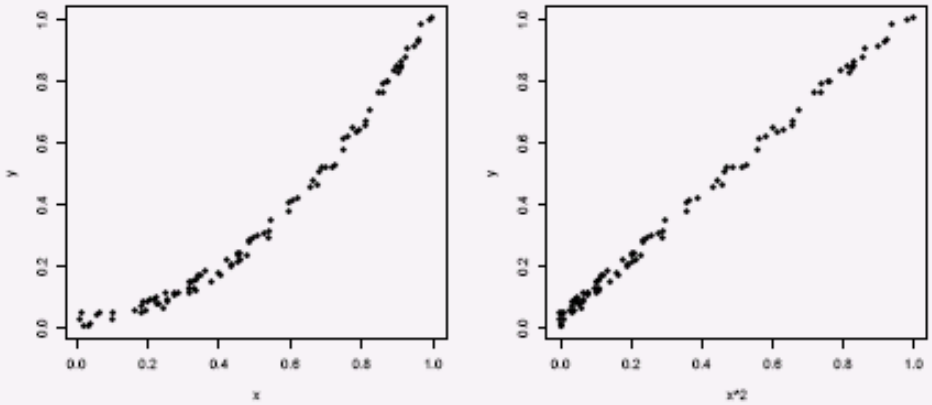
$$Y = \sum_{i=1}^{k} b_k z_k$$

$$Z_k = h_k(X)$$

Przykłady prostej transformacji

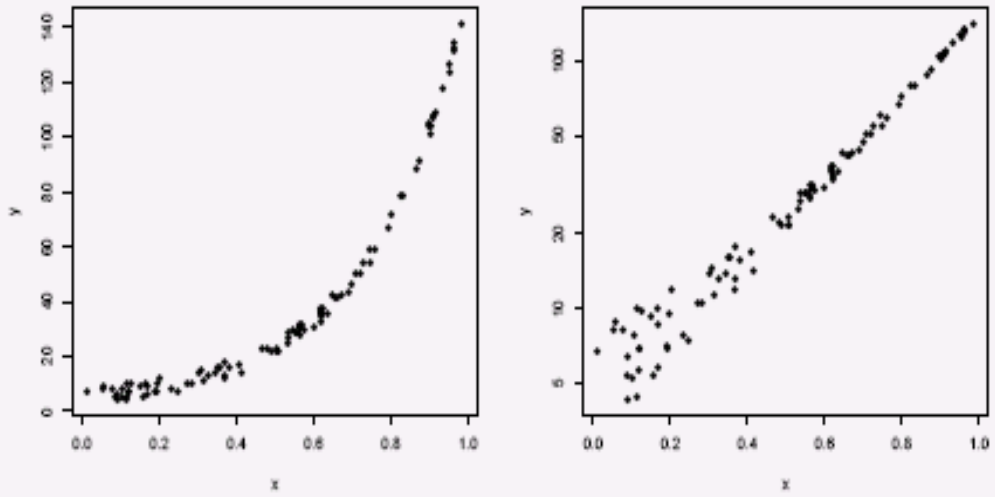
Co dają nam transformacje wielomianowe?

$$x' = x^2$$



logarytmowanie?

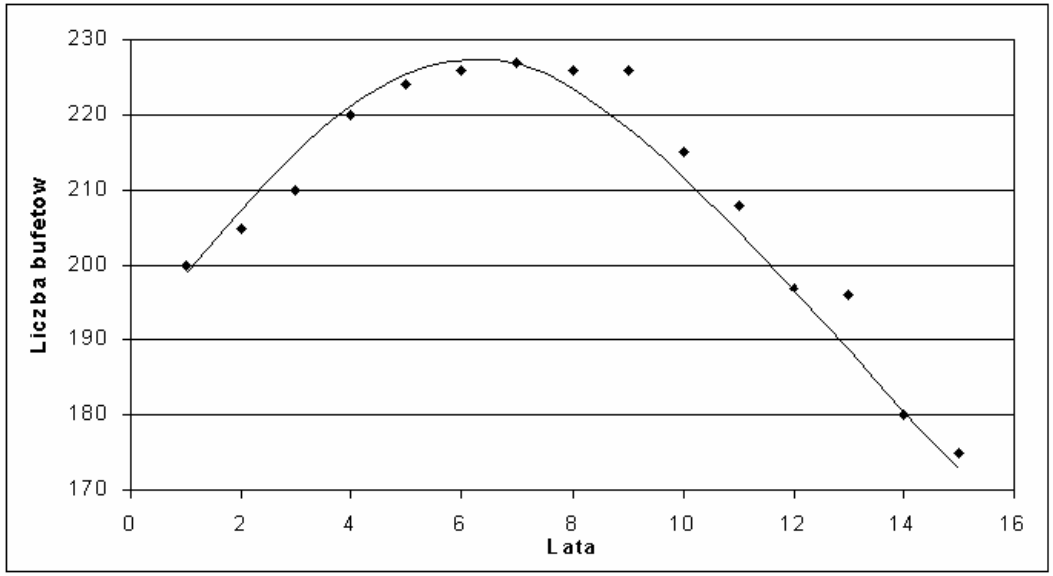
$$y' = log(y)$$



Przykład regresji nieliniowej

- Punkty żywieniowe w latach 1981-1995

Rok	Punkty	t
1981	200	1
1982	205	2
1983	210	3
1984	220	4
1985	224	5
1986	226	6
1987	227	7
1988	226	8
1989	226	9
1990	215	10
1991	208	11
1992	197	12
1993	196	13
1994	180	14
1995	175	15



Punkty żywieniowe c.d

Rok	y	Z1	Z2
1981	200	1	1
1982	205	2	4
1983	210	3	9
1984	220	4	16
1985	224	5	25
1986	226	6	36
1987	227	7	49
1988	226	8	64
1989	226	9	81
1990	215	10	100
1991	208	11	121
1992	197	12	144
1993	196	13	169
1994	180	14	196
1995	175	15	225

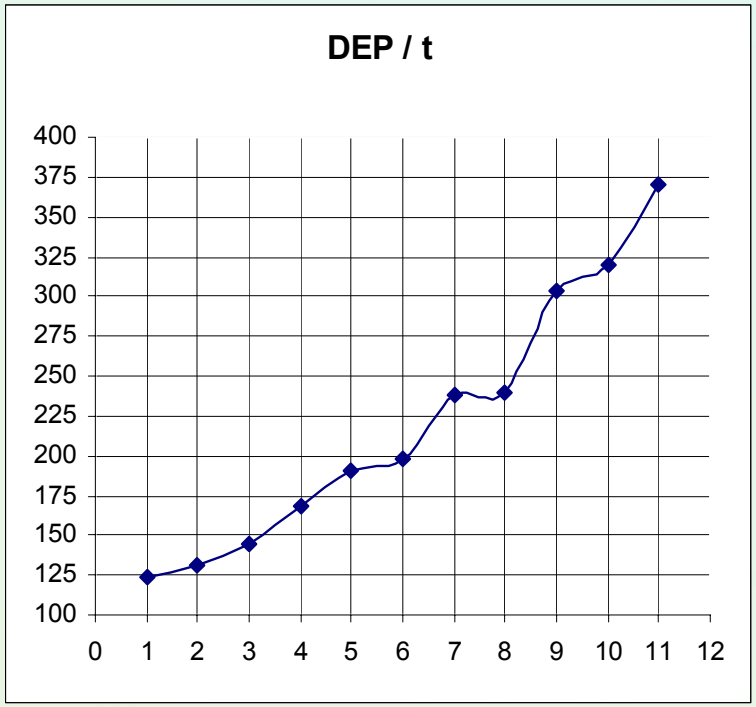
- Zakładamy, że kształt równania jest $y = a_0 + a_1 \cdot t + a_2 \cdot t^2$
- Wprowadzamy zmienne zastępcze $z_1 = t \quad z_2 = t^2$
- Rozwiązanie
  - a0=188
  - a1=11,031
  - a2=-0,814
- Weryfikacja
  - R2=0.996 s=3,37
  - Obie wartości statystyk *t < 0.05*

$$y = 188 + 11.031 \cdot t - 0.814 \cdot t^2$$

Przykład regresji nieliniowej – cz.2a

- Opisać kształtowania się depozytów złotowych w oddziale banku w kolejnych kwartałach lat 1994-1996

Kwartał	DEP	t
I 94	124	1
II 94	131	2
III 94	145	3
IV 94	169	4
I 95	190	5
II 95	198	6
III 95	238	7
IV 95	240	8
I 96	303	9
II 96	320	10
III 96	370	11

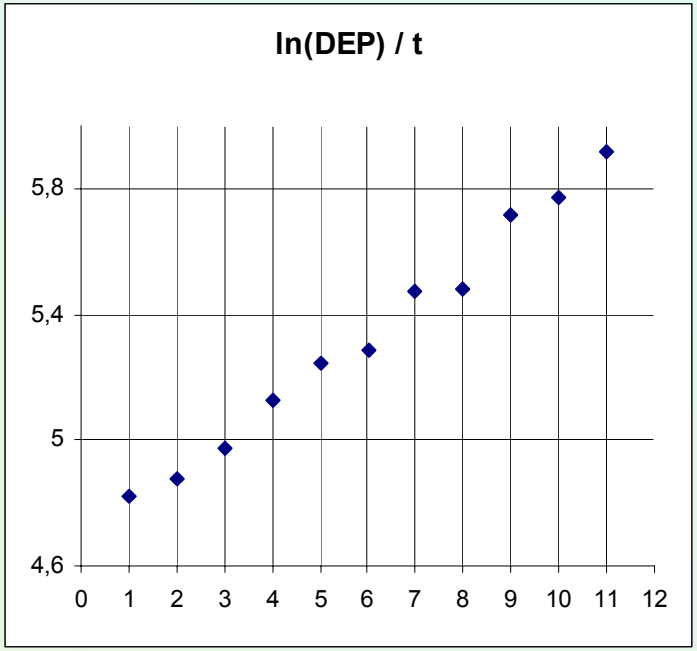


Hipoteza – wykładniczy przebieg $DEP = a \cdot e^{b \cdot t}$

Przykład regresji nieliniowej – cz.2b

- Opisać kształtowania się depozytów złotowych w oddziale banku w kolejnych kwartałach lat 1994-1996

t	*DEP*	Ln(*DEP*)
1	124	4.820
2	131	4,875
3	145	4,977
4	169	5,130
5	190	5,247
6	198	5,288
7	238	5,472
8	240	5,481
9	303	5,714
10	320	5,768
11	370	5,914



- Rozpatrujemy formę $\ln(DEP) = (\ln a) + b \cdot t$

Depozyty - rozwiązanie

- Rozwiązanie modelu przekształconego

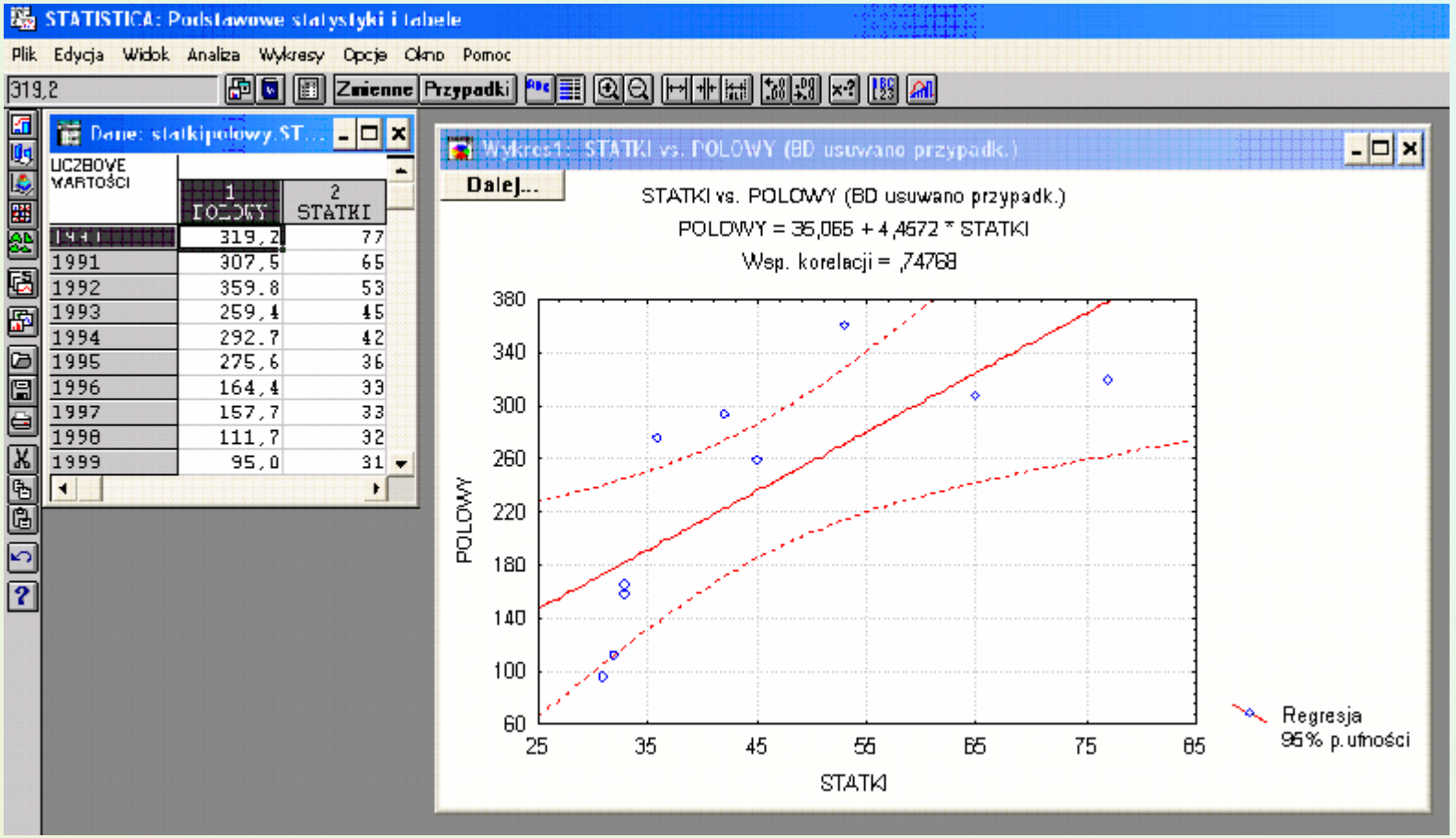
  $\ln(DEP)=4.671+0.111\cdot t$, R2=0.989, współczynniki istotne.

- Przekształcenie odwrotne

$$DEP = e^{4.671+0.111\cdot t} = 106.6 \cdot e^{0.111\cdot t}$$

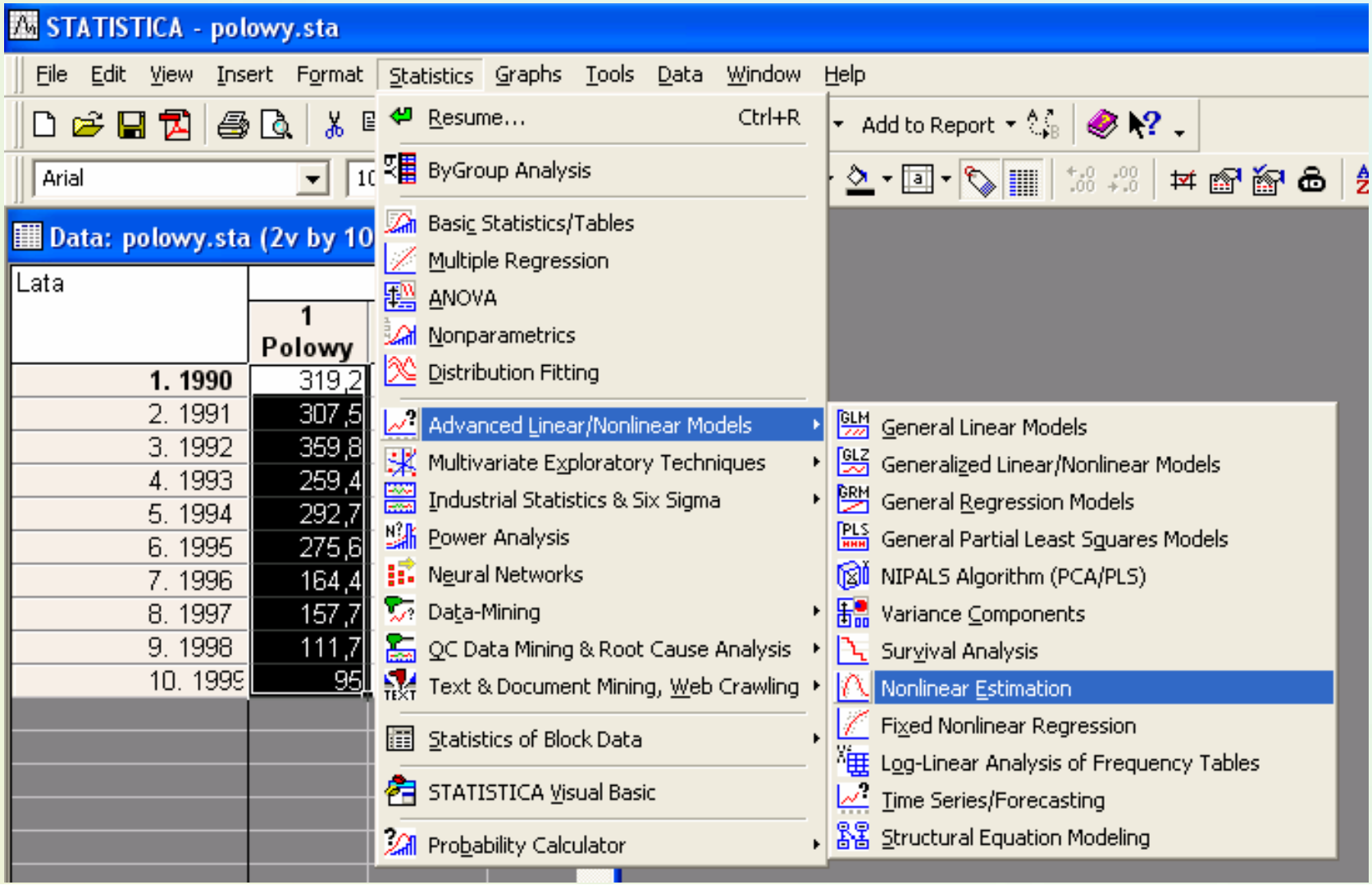
Inny przykład dla Statistica – patrz ćwiczenia laboratoryjne

- Dane nt. polskiego rybołówstwa dalekomorskiego (lata 90te).



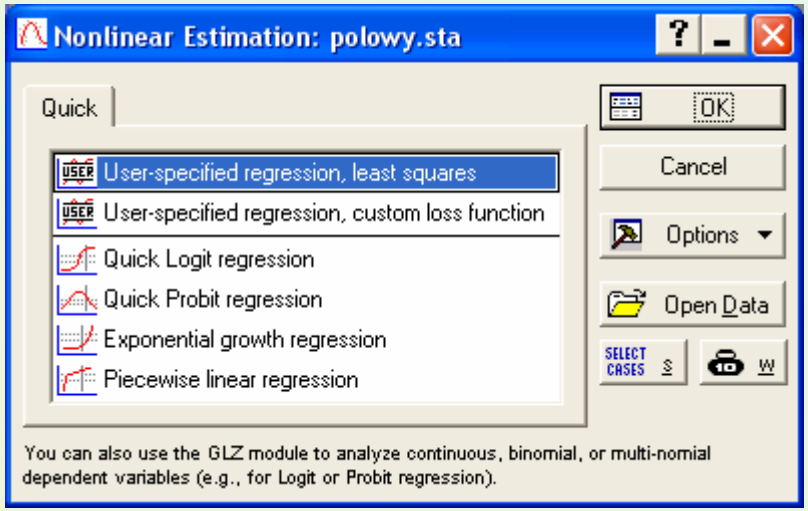
Statistica – poszukaj modeli nieliniowych

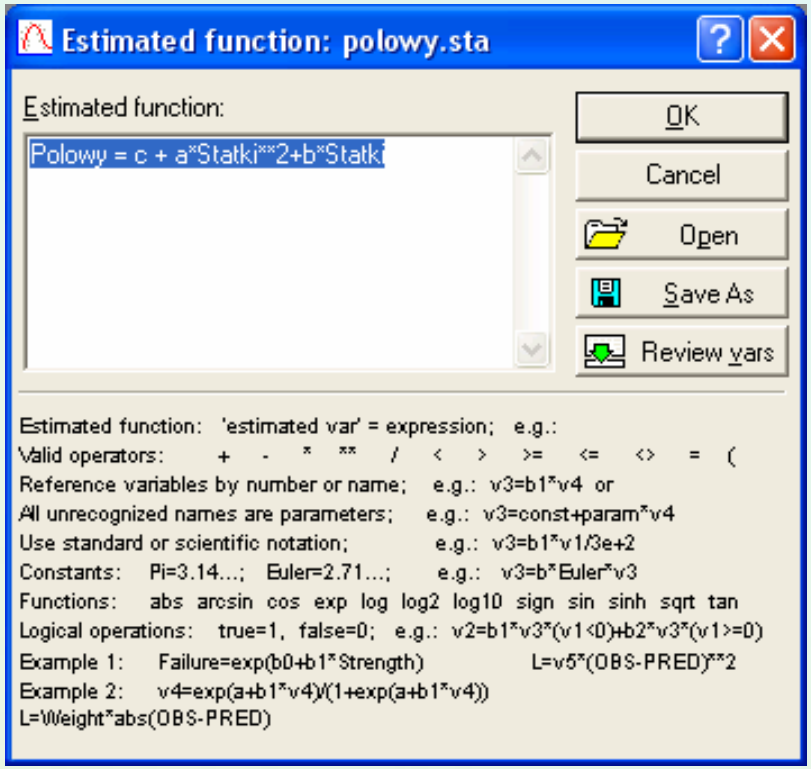
- Dwie opcje → na laboratorium sprawdź to?



Możliwości interakcji z użytkownikiem

- „User defined quadratic function”

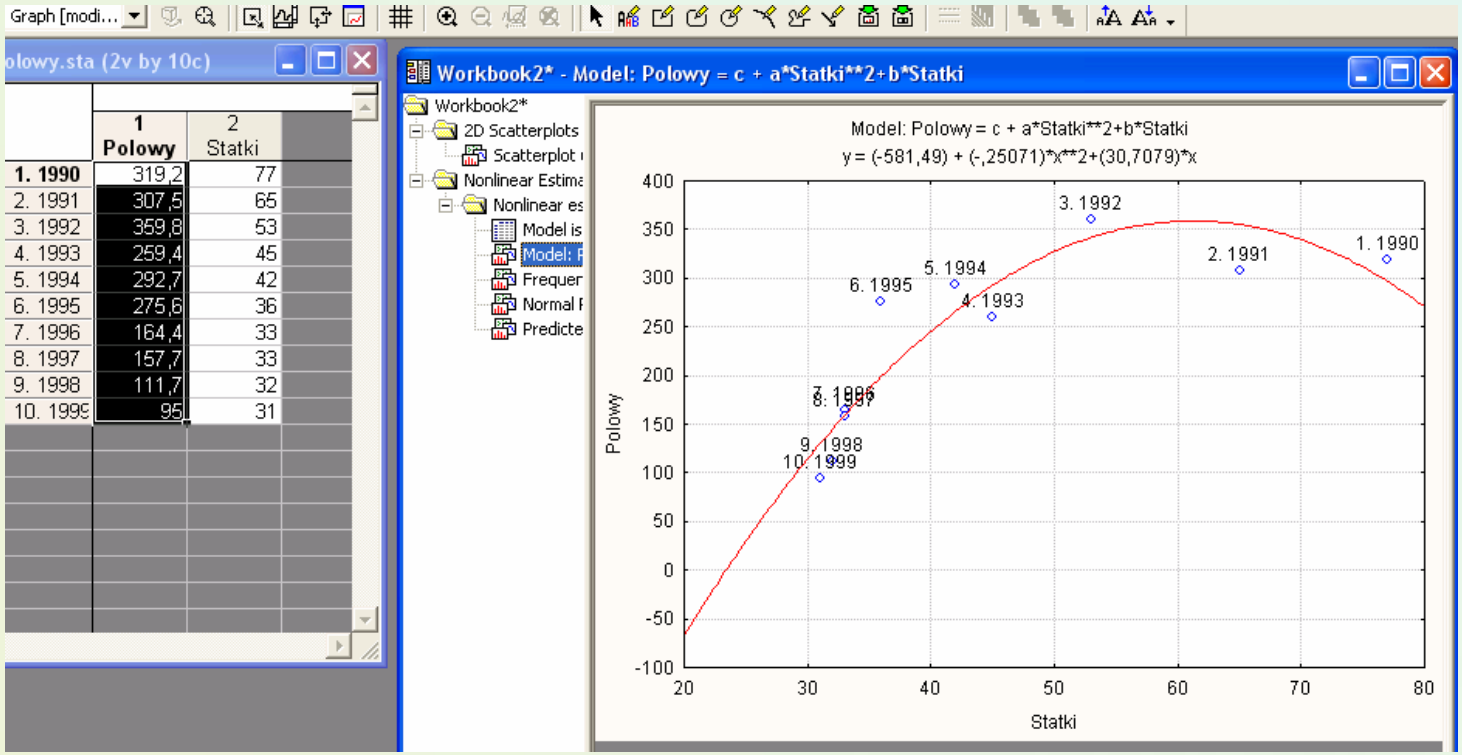




Regresja nieliniowa cd.

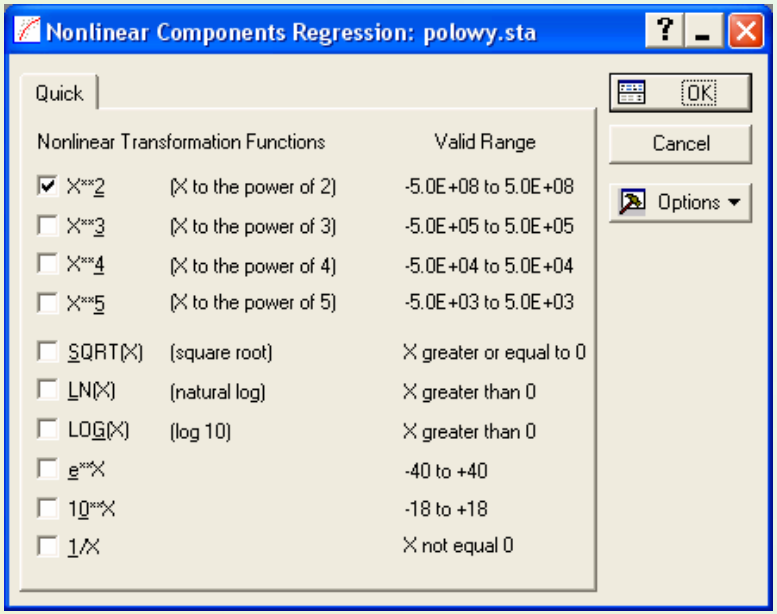
- Model funkcji kwadratowej (a co z innymi?)

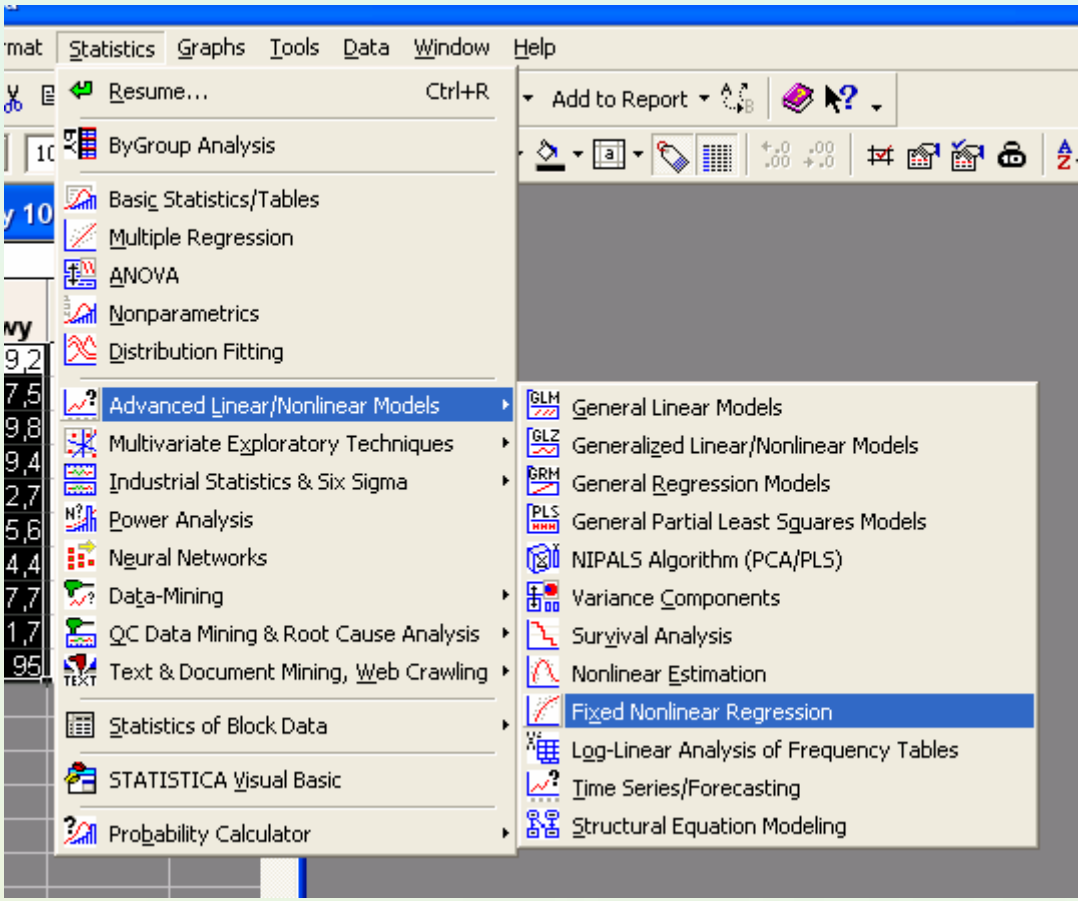
$$y = -0{,}25071 \cdot x^2 + 30{,}7079 \cdot x - 581{,}49$$



Statistica – inna opcja

- „Fitting fixed nonlinear basic functions”





Różne wskazówki co do transformacji

J.Koronacki, J.Mielniczuk: Statystyka rozdział 4.2

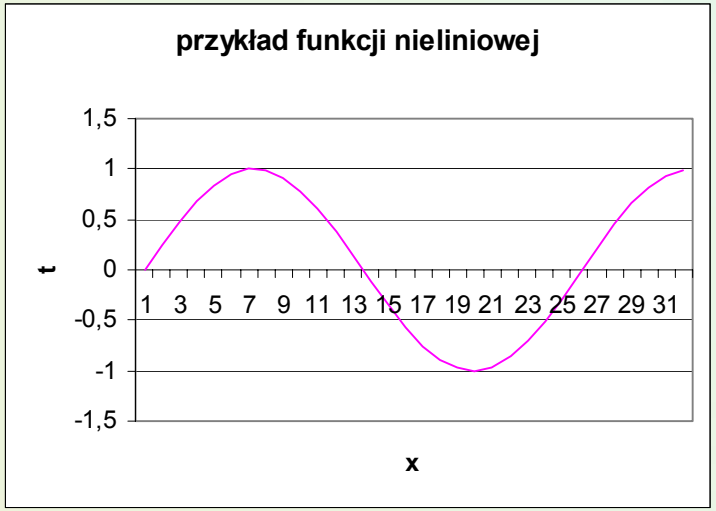
- Spójrz ogólne zasady doboru transformacji dla różnych typów zależności między zmiennymi.

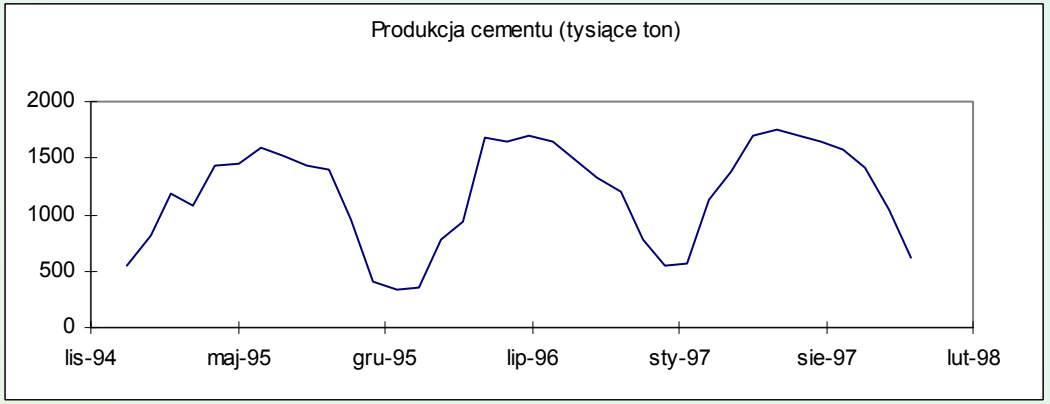
Także inne pozycje:

- M.Walesiak: Metody analizy danych marketingowych
- Rozdział 3.5 ogólne zasady transformacji liniowej dla typowych funkcji nieliniowych (dla kontekstu badań marketingowych)

Inne spojrzenie na nieliniowość

- Nie wszystkie modele są tak proste; funkcje nieliniowe mogą być bardziej złożone → nie nadają się do omówionych linearyzacji



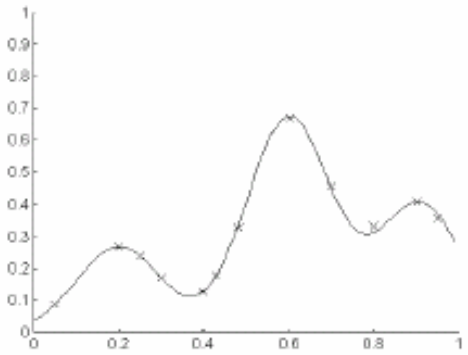


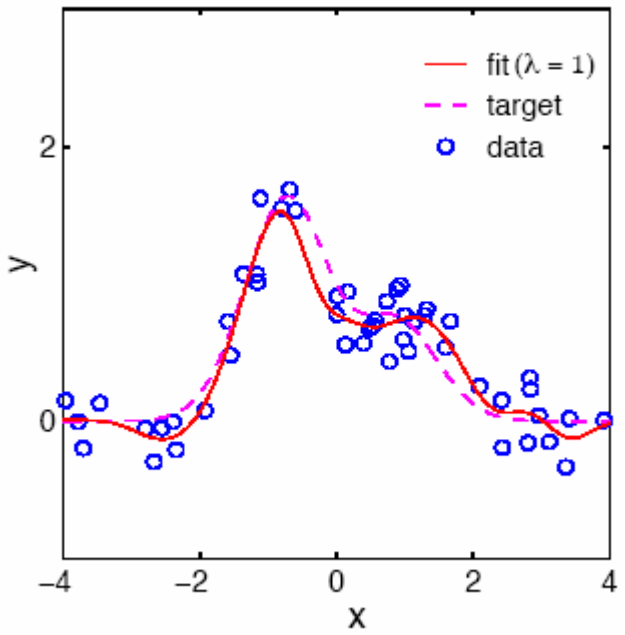
- Regresja nieparametryczna → przybliżona postać nie jest znana badaczowi z dokładnością do wybranych parametrów
- Przeczytaj więcej w rozdziale 5tym książki Koronacki, Ćwiek „Statystyczne systemy uczące się” wyd. 2

Różne funkcje

- Przykład trudniejszych funkcje – regresja nieparametryczna

data no	1	2	3	4	5	6	7	9	10	11	12	13
x	0.0500	0.2000	0.2500	0.3000	0.4000	0.4300	0.4800	0.6000	0.7000	0.8000	0.9000	0.9500
f(x)	0.0863	0.2662	0.2362	0.1687	0.1260	0.1756	0.3290	0.6694	0.4573	0.3320	0.4063	0.3535





Funkcje składane / sklejane

- Estymując funkcję regresji staramy się uwzględnić w modelu własności lokalne
- Składanie funkcji bazowych zdolnych lokalnie przybliżyć własności pewnych podobszarów dziedziny
- Regresyjne funkcje sklejane z węzłami

Locally weighted regression

$$y = \alpha + \sum_{j=1}^{p} f_j(\mathbf{x}, \beta)$$

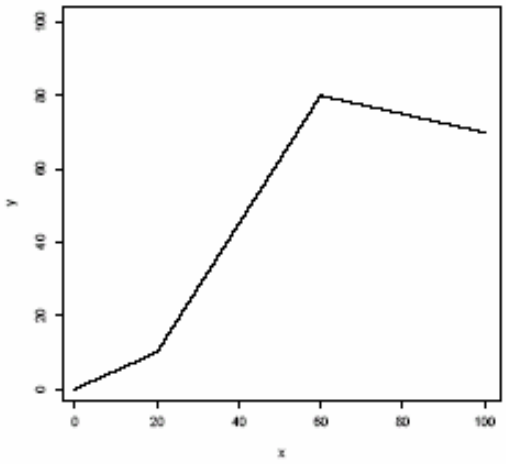
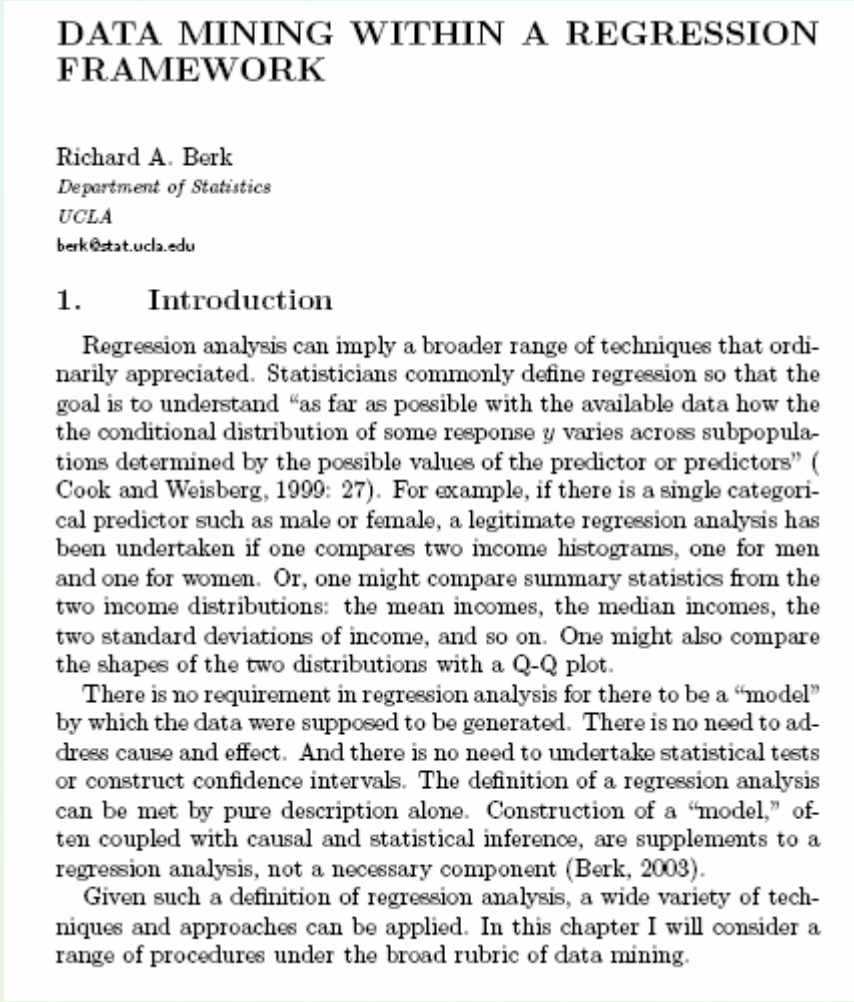


Figure 1.1. An Illustration of Linear Regression Splines with Two Knots

Zajrzyj do prac nt. zaawansowanych modeli

- Rozdział przeglądowy w „The Data Mining and Knowledge Discovery Handbook” O.Maimon, L. Rokach (eds), Springer 2005.
- Także J.Koronacki, Ćwik „Statystyczne systemy uczące” 2 wydanie → rozdział 5ty.



DATA MINING WITHIN A REGRESSION FRAMEWORK

Richard A. Berk
Department of Statistics
UCLA
berk@stat.ucla.edu

1. Introduction

Regression analysis can imply a broader range of techniques that ordinarily appreciated. Statisticians commonly define regression so that the goal is to understand “as far as possible with the available data how the the conditional distribution of some response y varies across subpopulations determined by the possible values of the predictor or predictors” (Cook and Weisberg, 1999: 27). For example, if there is a single categorical predictor such as male or female, a legitimate regression analysis has been undertaken if one compares two income histograms, one for men and one for women. Or, one might compare summary statistics from the two income distributions: the mean incomes, the median incomes, the two standard deviations of income, and so on. One might also compare the shapes of the two distributions with a Q-Q plot.

There is no requirement in regression analysis for there to be a “model” by which the data were supposed to be generated. There is no need to address cause and effect. And there is no need to undertake statistical tests or construct confidence intervals. The definition of a regression analysis can be met by pure description alone. Construction of a “model,” often coupled with causal and statistical inference, are supplements to a regression analysis, not a necessary component (Berk, 2003).

Given such a definition of regression analysis, a wide variety of techniques and approaches can be applied. In this chapter I will consider a range of procedures under the broad rubric of data mining.

Aproksymacja radialnymi funkcjami kołowymi RBF

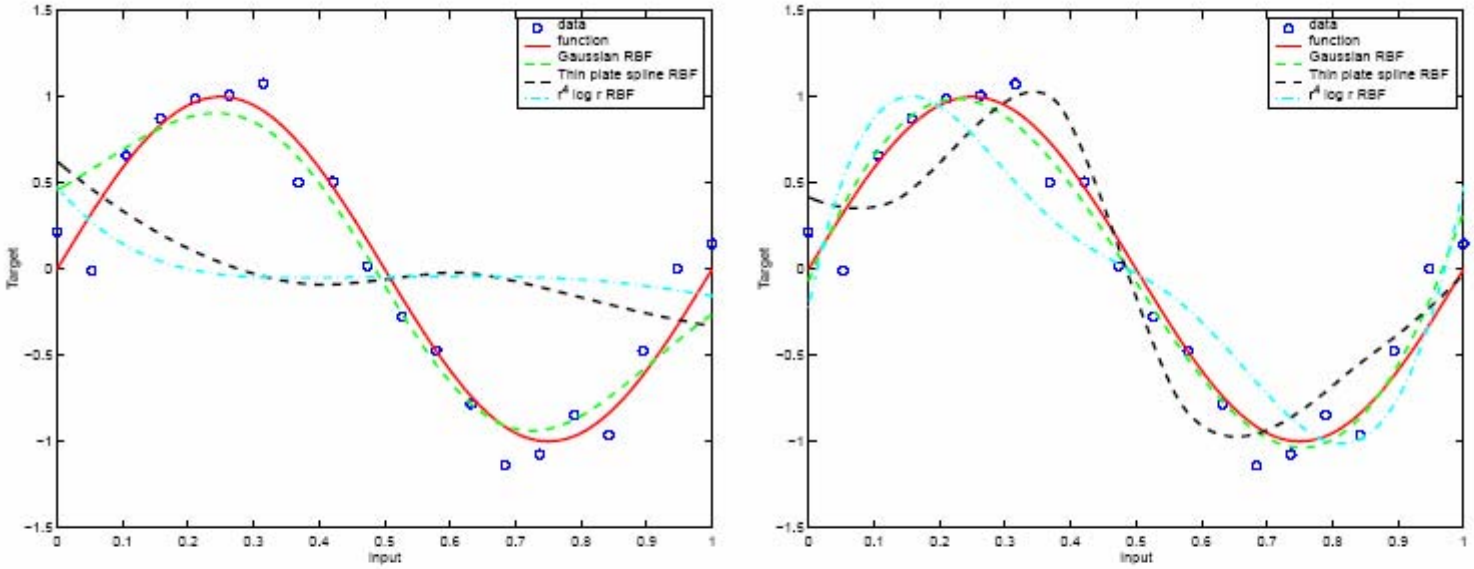
- Zadanie aproksymacji $f(\mathbf{x}_i) = d_i \quad \forall i = 1,\ldots,N$
- Przyjmijmy funkcje liniową względem parametrów w wykorzystującą funkcje o symetrii kołowej RBF

$$f(\mathbf{x}) = \sum_{i=1}^{m} w_i \cdot \varphi(\|\mathbf{x} - c_i\|)$$

- Radialna funkcja bazowa $\rightarrow$ funkcja φ o postaci $\varphi(\mathbf{x},\mathbf{c}) = \varphi(r(\mathbf{x},\mathbf{c}))$, gdzie r jest odległością między punktami **x** i **c**. Punkt c nazywamy centrum
- Związek funkcji radialnym z funkcjami jądrowymi (kernals), z parametrem σ szerokością jądra

Przykład aproksymacji funkcji sinsus

- Za wykład prof. A.Bartkowiak Uniw. Wrocławski



Rysunek 9.3: *Aproksymacja funkcji sinus wykonana za pomocą sieci* RBF *o H=2 (lewa) i H=4 (prawa) neuronach. Sieci zostały wytrenowane na podstawie 20-elementowej zaburzonej próbki wylosowanej z sinusoidy. Pliki rr2c.eps i rr4c.eps*

Metody doboru zmiennych do modelu

- Zmienne wybiera się na podstawie wiedzy dziedzinowej.
- Wymagania nt. własności zmiennych niezależnych:
  - Są silnie skorelowanych ze zmienną, którą objaśniają.
  - Są nieskorelowane lub co najwyżej słabo skorelowane ze sobą.
  - Charakteryzują się dużą zmiennością.
- Jak wykorzystać współczynniki korelacji?

$$r^* = \sqrt{\frac{t^2_{\alpha,n-2}}{n-2+t^2_{\alpha,n-2}}}$$

Ocena zmiennych objaśniających

- Przykład doboru zmiennych do modelu opisującego miesięczne spożycie ryb (w kg na osobę) w zależności od: spożycia mięsa x_1, warzyw x_2, owoców x_3, tłuszczów x_4 oraz wydatków na lekarstwa x_5.

nr	y	X1	X2	X3	X4	x5
1	3	3	0,63	0,63	0,12	14,1
2	3	3	1,07	1,07	0,14	12,77
3	3	3	0,44	0,44	0,1	11
4	3	2	0,26	0,26	0,04	44
5	0	0	0,01	0,0	0,0	60
6	0	0	0,02	0,01	0,0	66
7	0	0	0,02	0,01	0,01	53
8	5	4	0,09	0,09	0,03	60
9	4	2	0,56	0,56	0,19	3
10	3	2	0,11	0,11	0,05	3
11	7	7	1,46	1,46	0,34	23
12	5	5	1,22	1,22	0.24	30
13	5	5	1,22	1,22	0,26	30
14	2	1	0,31	0,13	0,05	39
15	3	2	0,4	0,19	0,05	56

Dobór zmiennych do modelu

- Współczynniki zmienności

y	x1	x2	x3	x4	X5
0,635	0,754	0,917	1,0	0,944	0,632

- Macierz współczynników korelacji

	y	x1	x2	x3	x4	X5
y	1					
x1	0,950	1				
x2	0,750	0,843	1			
x3	0,748	0,851	0,991	1		
x4	0,813	0,860	0,946	0,951	1	
x5	-0,442	-0,395	-0,477	-0,503	-0,539	1

Trochę obliczeń

- Wartość krytyczna $r^* = \sqrt{\frac{4{,}6656}{13 + 4{,}6656}} = \sqrt{0.264107} = 0.5139$
- Słaba korelacja?

  r(y,x5) =-0.442 $\rightarrow$ odrzucamy x5
- Wybieramy najsilniejszą zmienną

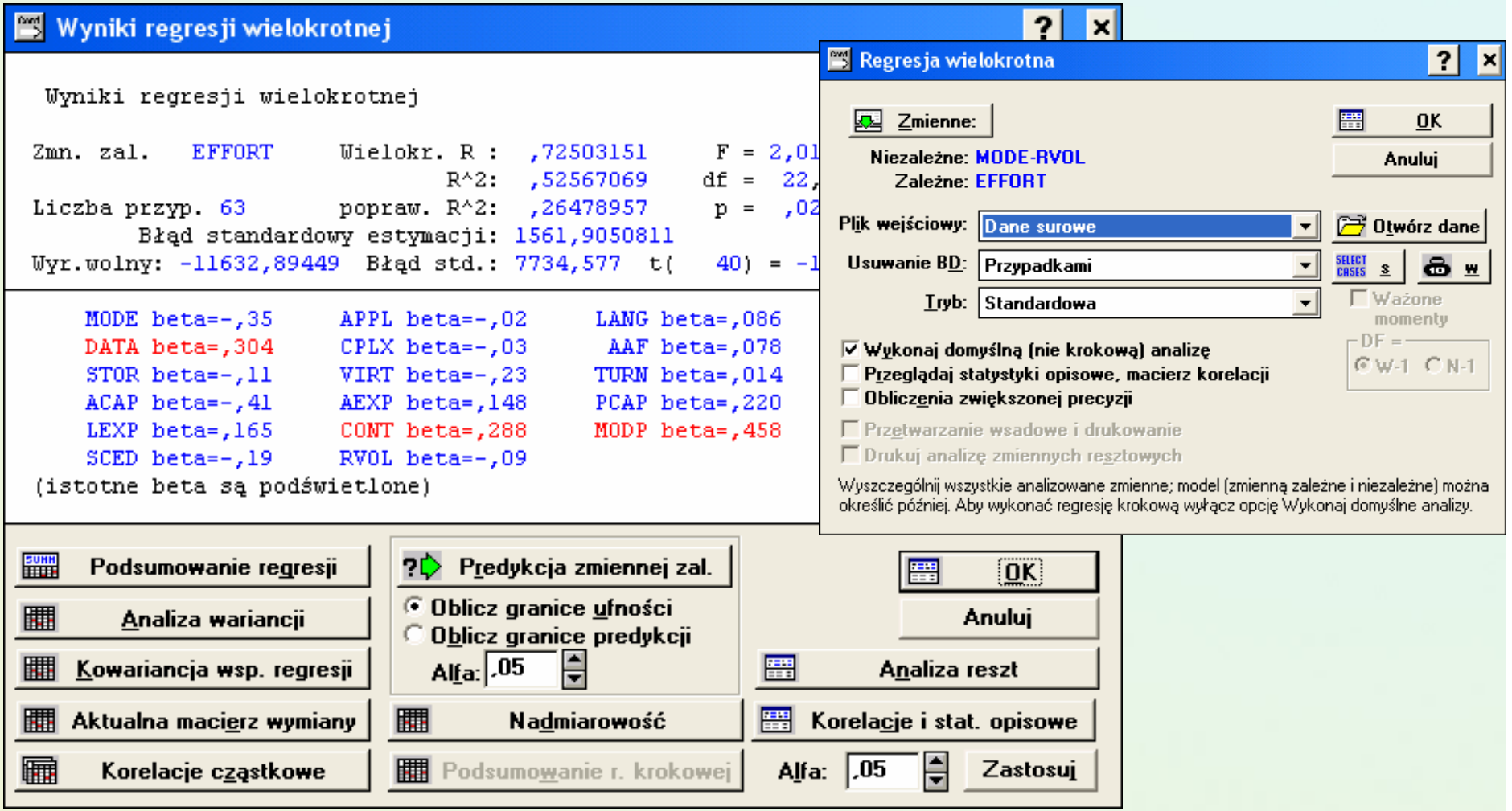
  r(y,x1)=r1=0.950 $\rightarrow$ wybieramy x1

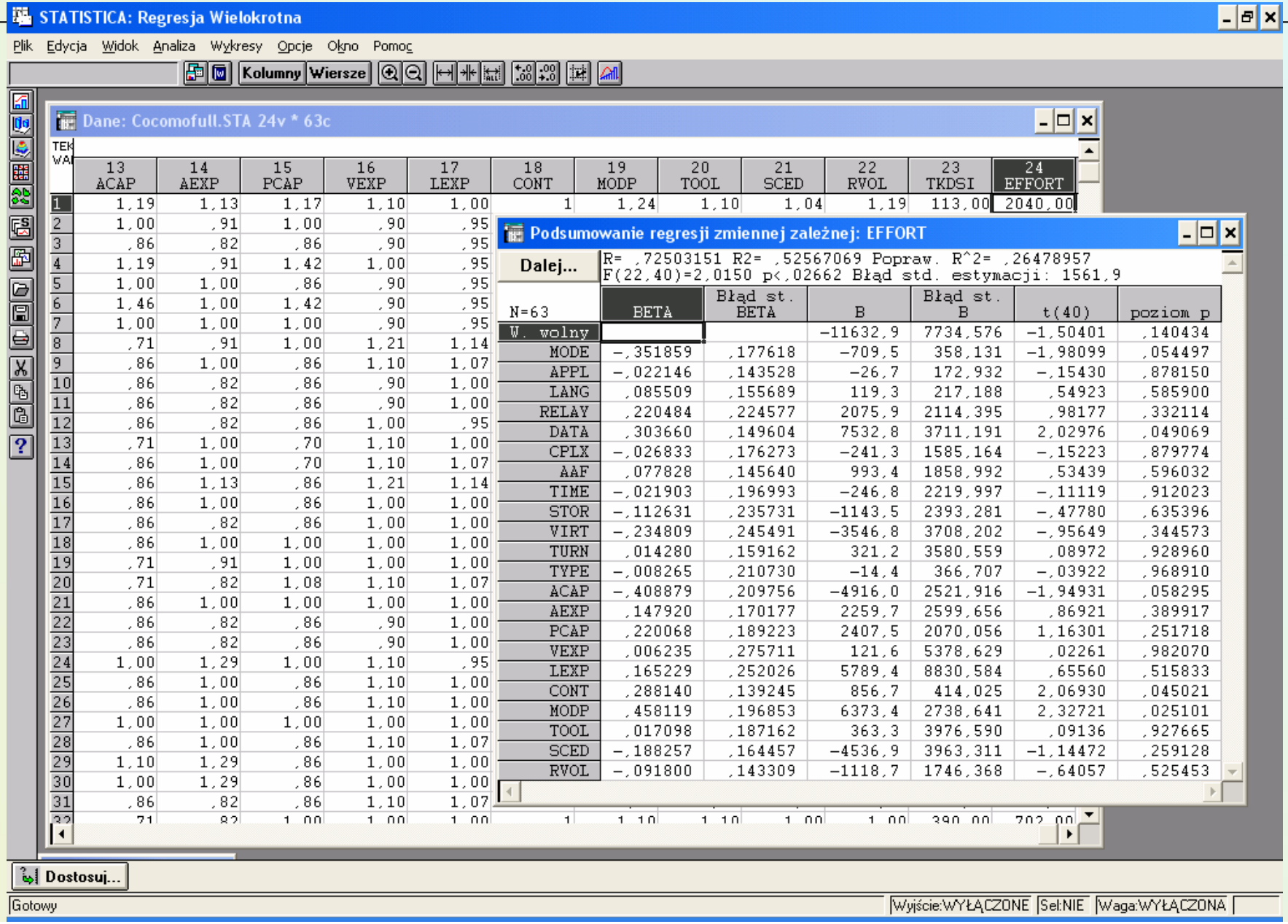
Co z pozostałymi zmiennymi?

Regresja krokowa

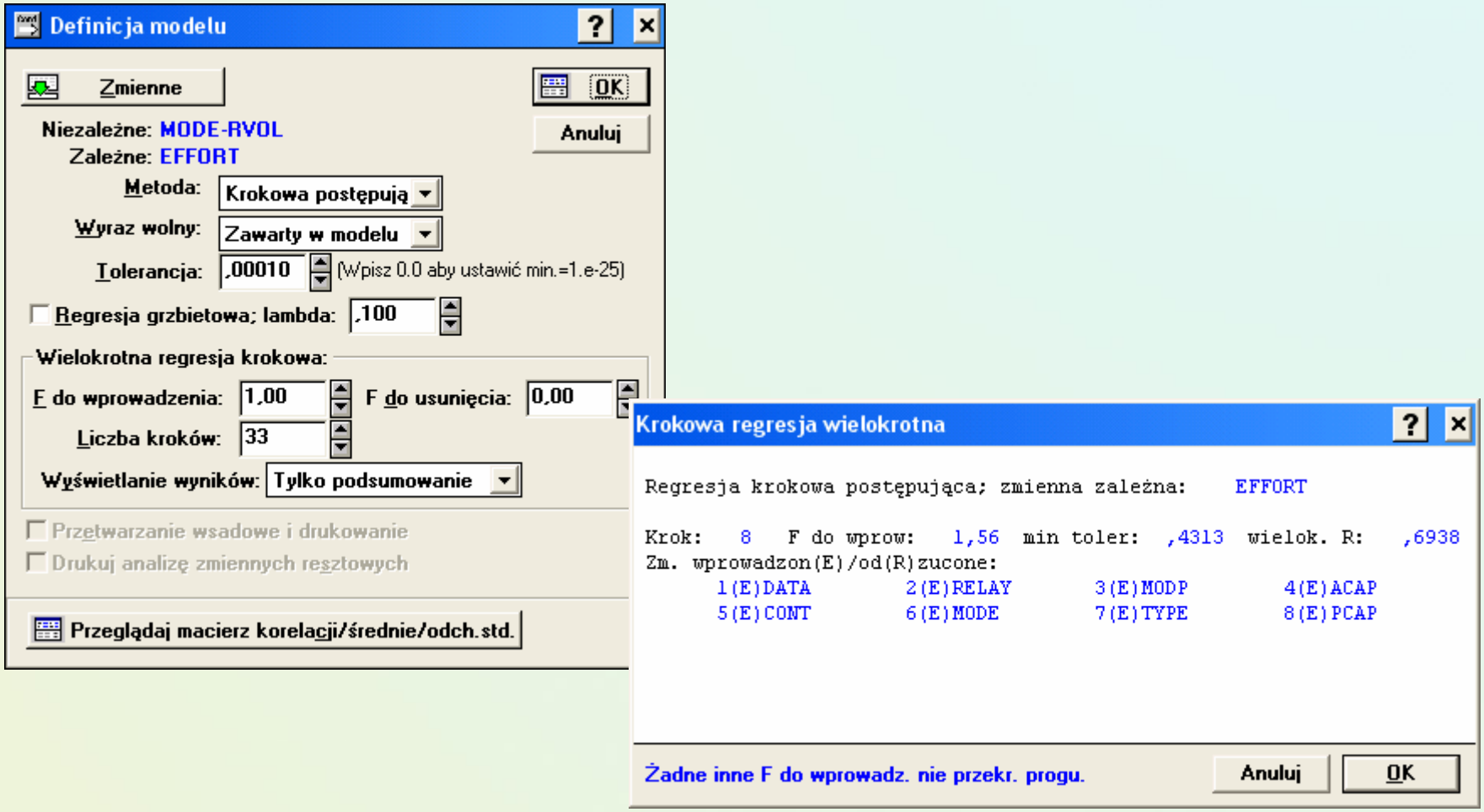
- Postępująca (*forward*)
  - Zakłada kolejne dołączanie do listy zmiennych objaśniających tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.
- Wsteczna (*backward*)
  - Usuwamy ze zbioru zmiennych, ta które mają najmniejszy wpływ na zmienną zależną.
  - Stosując R2 lub testy istotności współczynników modelu (*F*).

Regresja wielokrotna - Statistica




STATISTICA: Regresja Wielokrotna
Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc
Kolumny Wiersze
Dane: Cocomofull.STA 24v * 63c
13 ACAP 14 AEXP 15 PCAP 16 VEXP 17 LEXP 18 CONT 19 MODP 20 TOOL 21 SCED 22 RVOL 23 TKDSI 24 EFFORT
1 1,19 1,13 1,17 1,10 1,00 1 1,24 1,10 1,04 1,19 113,00 2040,00
2 1,00 ,91 1,00 ,90 ,95
3 ,86 ,82 ,86 ,90 ,95
4 1,19 ,91 1,42 1,00 ,95
5 1,00 1,00 ,86 ,90 ,95
6 1,46 1,00 1,42 ,90 ,95
7 1,00 1,00 1,00 ,90 ,95
8 ,71 ,91 1,00 1,21 1,14
9 ,86 1,00 ,86 1,10 1,07
10 ,86 ,82 ,86 ,90 1,00
11 ,86 ,82 ,86 ,90 1,00
12 ,86 ,82 ,86 1,00 ,95
13 ,71 1,00 ,70 1,10 1,00
14 ,86 1,00 ,70 1,10 1,07
15 ,86 1,13 ,86 1,21 1,14
16 ,86 1,00 ,86 1,00 1,00
17 ,86 ,82 ,86 1,00 1,00
18 ,86 1,00 1,00 1,00 1,00
19 ,71 ,91 1,00 1,00 1,00
20 ,71 ,82 1,08 1,10 1,07
21 ,86 1,00 1,00 1,00 1,00
22 ,86 ,82 ,86 ,90 1,00
23 ,86 ,82 ,86 ,90 1,00
24 1,00 1,29 1,00 1,10 ,95
25 ,86 1,00 ,86 1,10 1,00
26 ,86 1,00 ,86 1,10 1,00
27 1,00 1,00 1,00 1,00 1,00
28 ,86 1,00 ,86 1,10 1,07
29 1,10 1,29 ,86 1,00 1,00
30 1,00 1,29 ,86 1,00 1,00
31 ,86 ,82 ,86 1,10 1,07
Podsumowanie regresji zmiennej zależnej: EFFORT
Dalej...
R= ,72503151 R2= ,52567069 Popraw. R^2= ,26478957
F(22,40)=2,0150 p<,02662 Błąd std. estymacji: 1561,9
N=63 BETA Błąd st. BETA B Błąd st. B t(40) poziom p
W. wolny -11632,9 7734,576 -1,50401 ,140434
MODE -,351859 ,177618 -709,5 358,131 -1,98099 ,054497
APPL -,022146 ,143528 -26,7 172,932 -,15430 ,878150
LANG ,085509 ,155689 119,3 217,188 ,54923 ,585900
RELAY ,220484 ,224577 2075,9 2114,395 ,98177 ,332114
DATA ,303660 ,149604 7532,8 3711,191 2,02976 ,049069
CPLX -,026833 ,176273 -241,3 1585,164 -,15223 ,879774
AAF ,077828 ,145640 993,4 1858,992 ,53439 ,596032
TIME -,021903 ,196993 -246,8 2219,997 -,11119 ,912023
STOR -,112631 ,235731 -1143,5 2393,281 -,47780 ,635396
VIRT -,234809 ,245491 -3546,8 3708,202 -,95649 ,344573
TURN ,014280 ,159162 321,2 3580,559 ,08972 ,928960
TYPE -,008265 ,210730 -14,4 366,707 -,03922 ,968910
ACAP -,408879 ,209756 -4916,0 2521,916 -1,94931 ,058295
AEXP ,147920 ,170177 2259,7 2599,656 ,86921 ,389917
PCAP ,220068 ,189223 2407,5 2070,056 1,16301 ,251718
VEXP ,006235 ,275711 121,6 5378,629 ,02261 ,982070
LEXP ,165229 ,252026 5789,4 8830,584 ,65560 ,515833
CONT ,288140 ,139245 856,7 414,025 2,06930 ,045021
MODP ,458119 ,196853 6373,4 2738,641 2,32721 ,025101
TOOL ,017098 ,187162 363,3 3976,590 ,09136 ,927665
SCED -,188257 ,164457 -4536,9 3963,311 -1,14472 ,259128
RVOL -,091800 ,143309 -1118,7 1746,368 -,64057 ,525453
Dostosuj...
Gotowy
Wyjście:WYŁĄCZONE Sel:NIE Waga:WYŁĄCZONA

Regresja krokowa



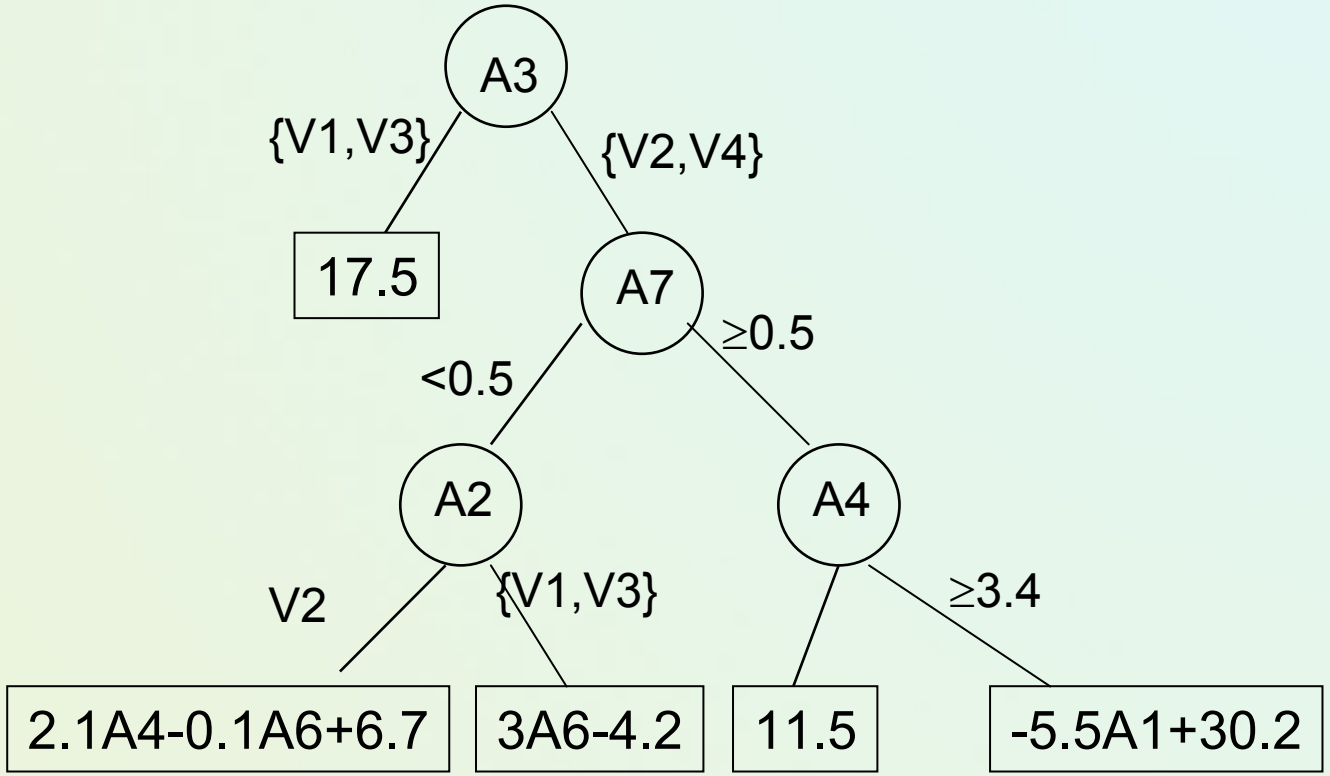
Inne zaawansowane modele regresji

- Generalized linear model:
  - Foundation on which linear regression can be applied to modeling categorical response variables
  - Variance of y is a function of the mean value of y, not a constant
  - **Logistic regression**: models the prob. of some event occurring as a linear function of a set of predictor variables
  - Poisson regression: models the data that exhibit a Poisson distribution
- Log-linear models: (for categorical data)
  - Approximate discrete multidimensional prob. distributions
  - Also useful for data compression and smoothing
- Regression trees and model trees
  - Trees to predict continuous values rather than class labels

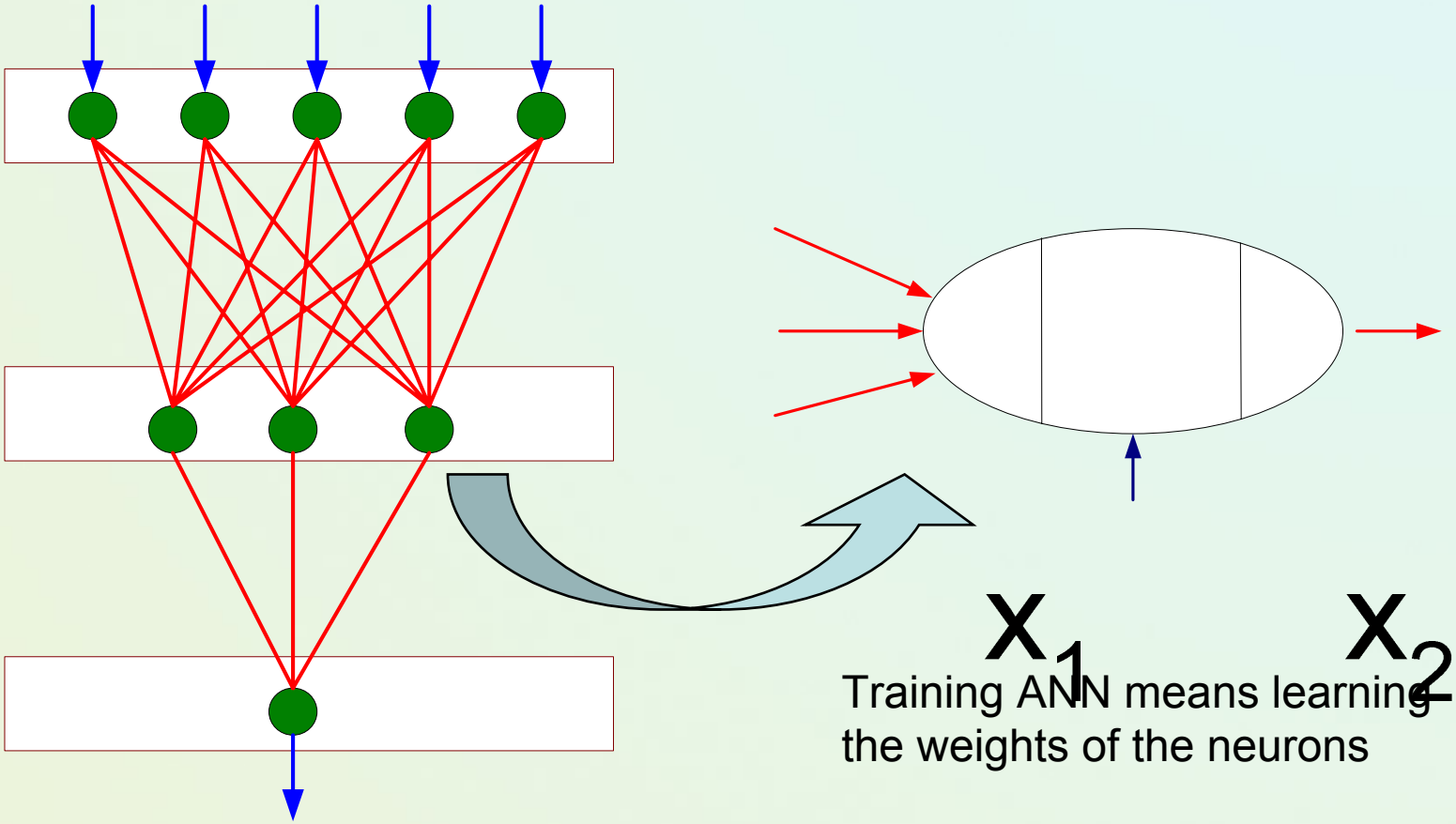
Regression Trees and Model Trees

- Regression tree: proposed in **CART** system (Breiman et al. 1984)
  - CART: Classification And Regression Trees
  - Each leaf stores a *continuous-valued prediction*
  - It is the *average value of the predicted attribute* for the training tuples that reach the leaf
- Model tree: proposed by Quinlan (1992)
  - Each leaf holds a regression model—a multivariate linear equation for the predicted attribute
  - A more general case than regression tree
- Regression and model trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model

An example regression tree



Sztuczne sieci neuronowe - predykcja



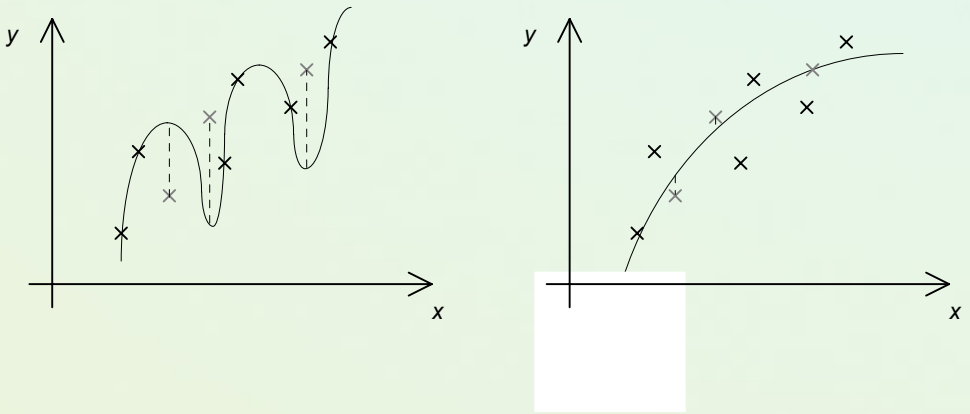
x_1 x_2 x

Training ANN means learning the weights of the neurons

Input Layer

Stosowanie sieci neuronowych do predykcji

- Wybór rodzaju sieci: warstwowa MLP albo RBF
- MLP
  - Topologia sieci (ile neuronów w warstwie ukrytej)
  - Rodzaj neuronów (dobór funkcji aktywacji i parametrów)
  - Algorytm uczenia (BP) i warunki zatrzymania
  - Unikaj przeuczenia!
  - Konieczność posiadania właściwego i odpowiednio licznego zbioru uczącego.



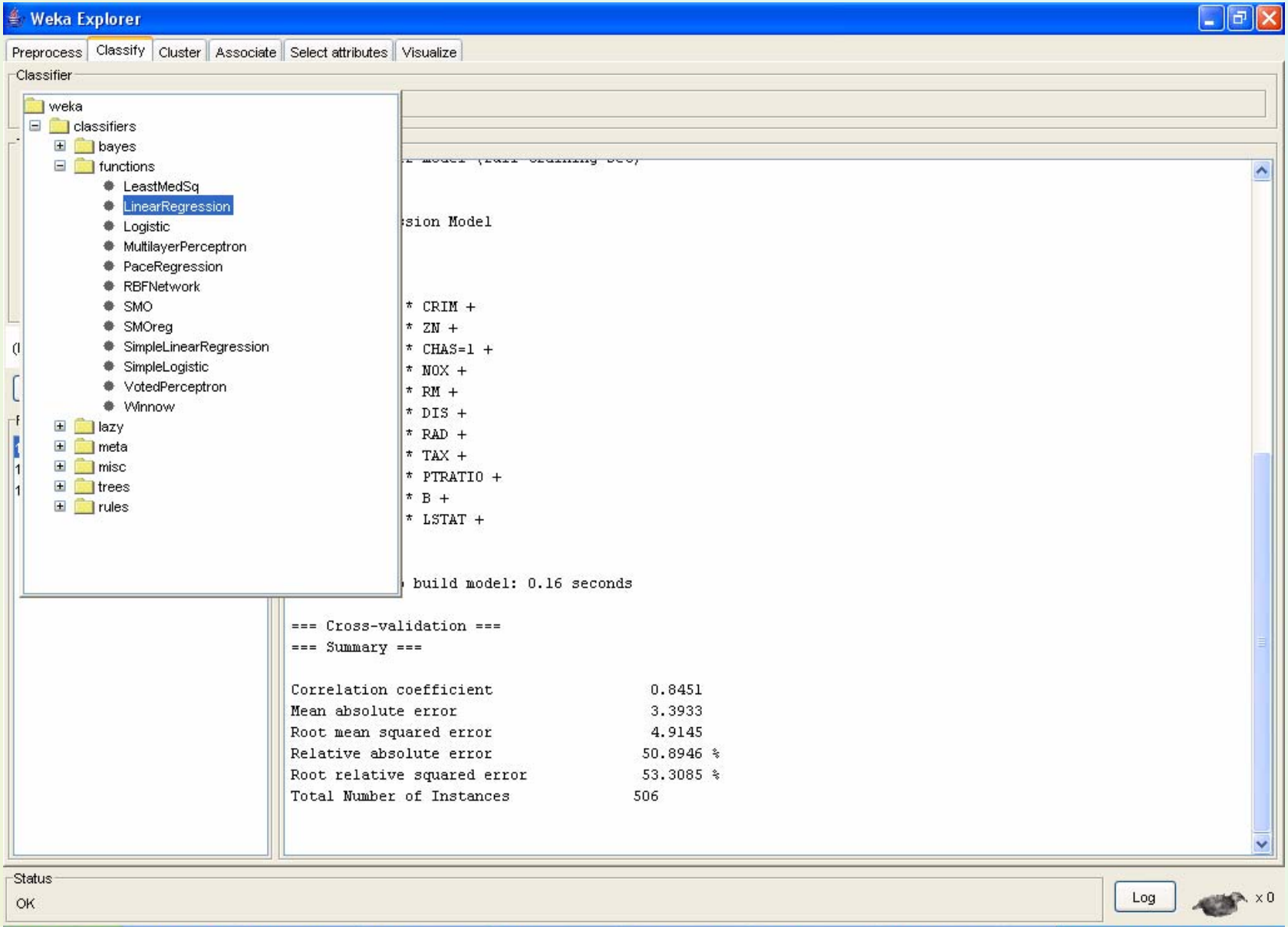
Logistic regression

- Problem: some assumptions violated when linear regression is applied to classification problems
- *Logistic* regression: alternative to linear regression
  - Designed for classification problems
  - Tries to estimate class probabilities directly
    - Does this using the *maximum likelihood* method
  - Uses this linear model:

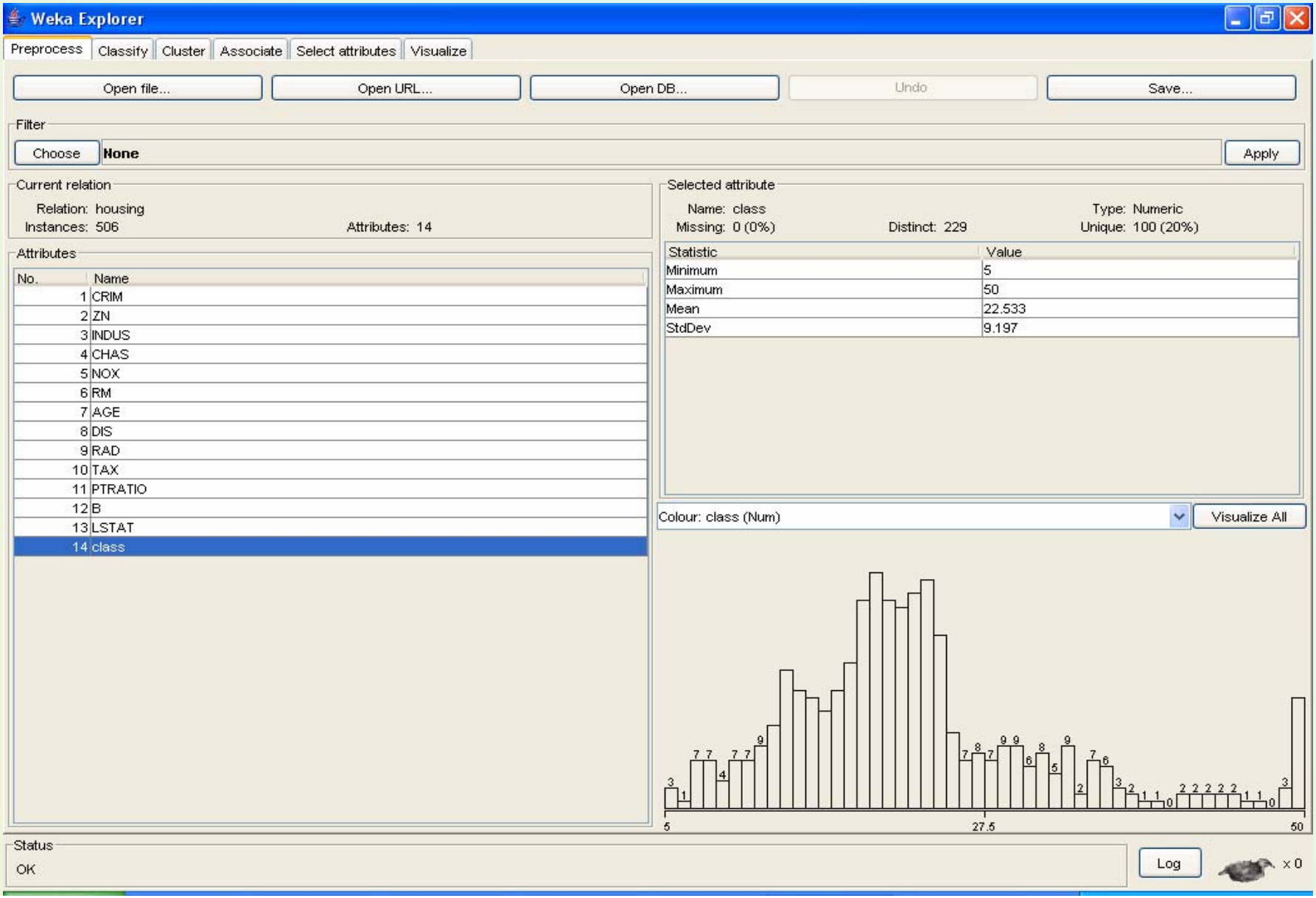
$$\log\left(\frac{P}{1-P}\right) = w_0 a_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

P= Class probability

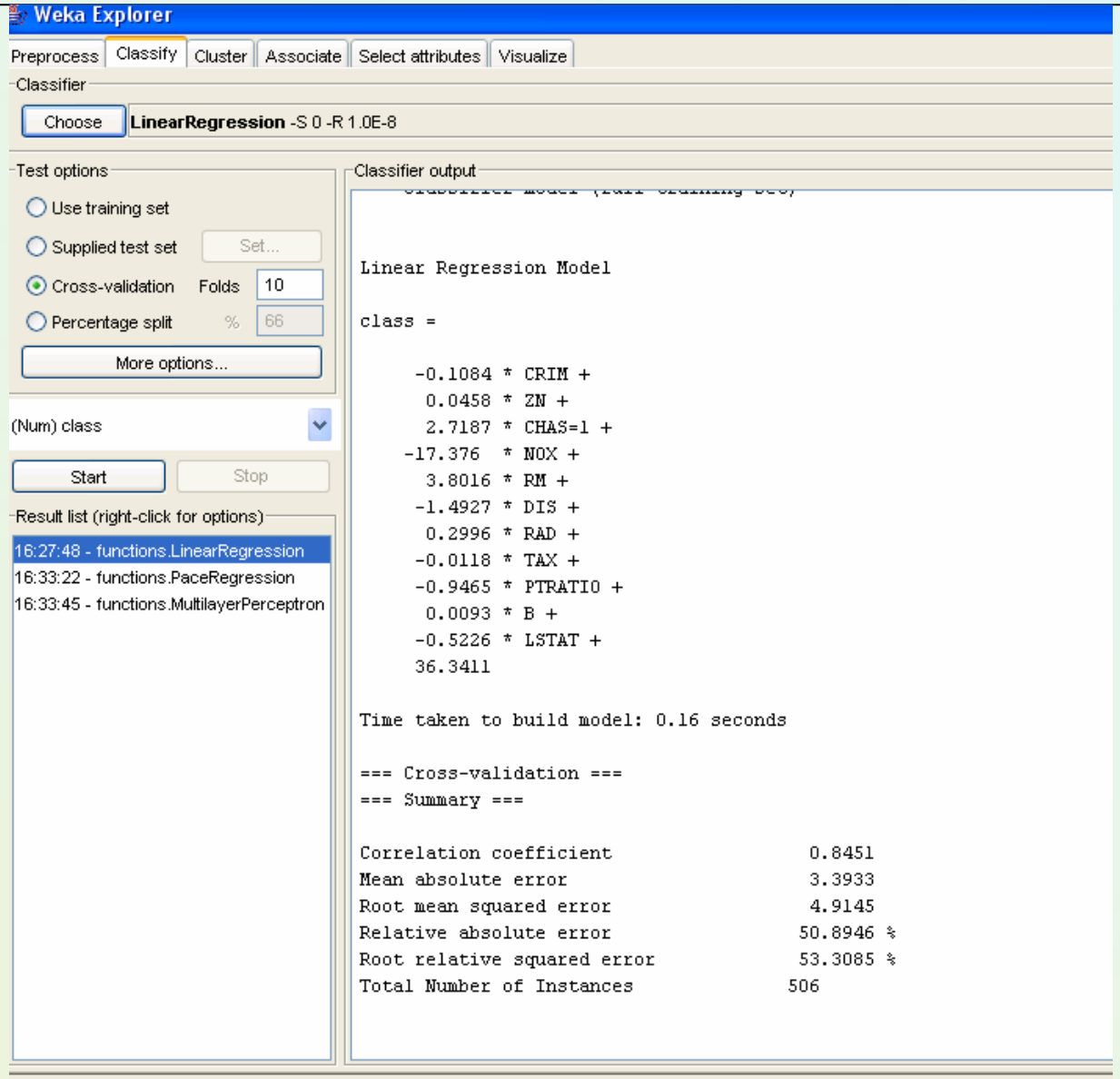
Modele predykcji zmiennej liczbowej w WEKA



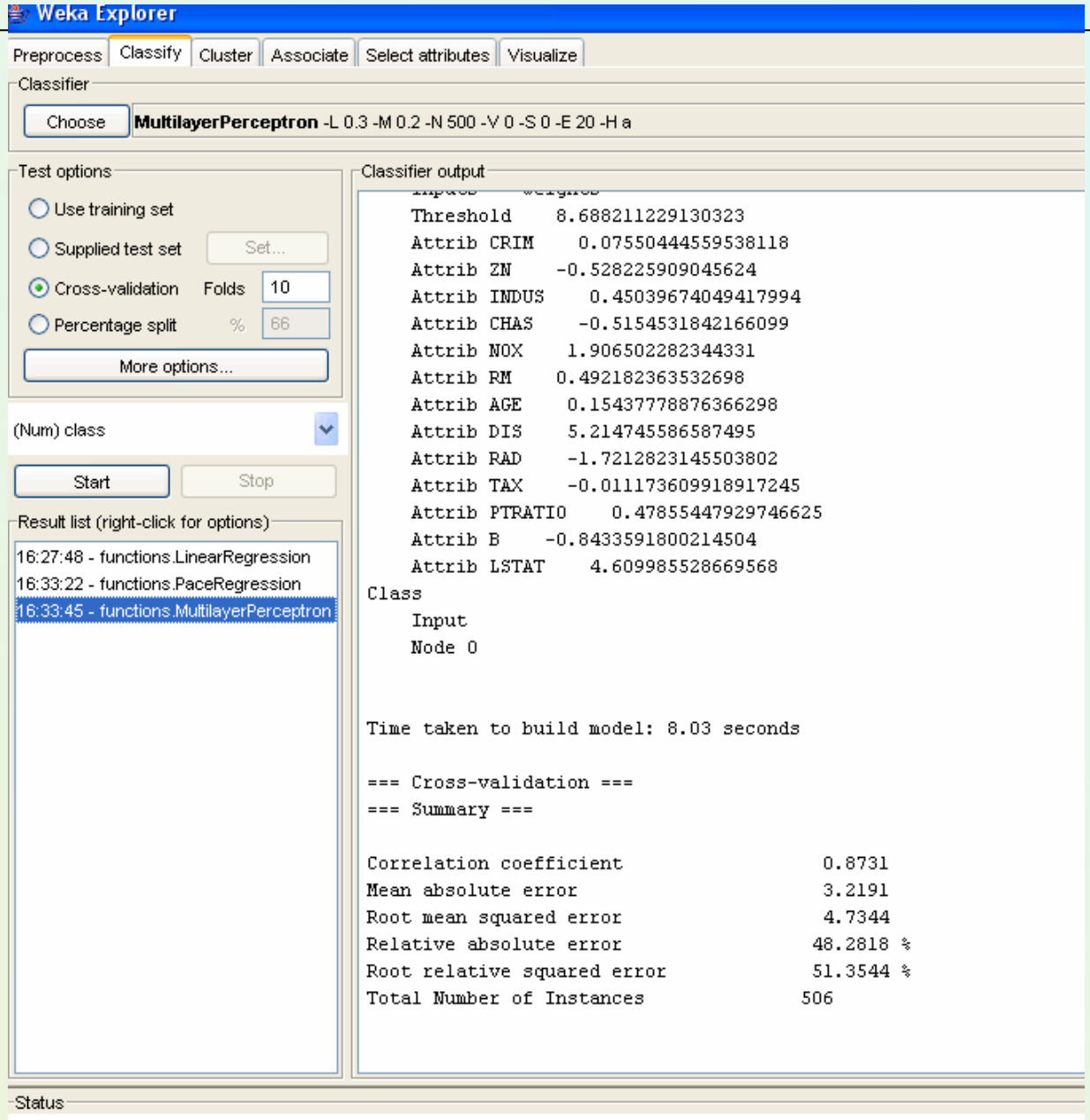
Wczytywanie danych



Regresja wielokrotna liniowa



Inne metody


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) class
Start
Stop
Result list (right-click for options)
16:27:48 - functions.LinearRegression
16:33:22 - functions.PaceRegression
16:33:45 - functions.MultilayerPerceptron
Classifier output
Threshold 8.688211229130323
Attrib CRIM 0.07550444559538118
Attrib ZN -0.528225909045624
Attrib INDUS 0.45039674049417994
Attrib CHAS -0.5154531842166099
Attrib NOX 1.906502282344331
Attrib RM 0.492182363532698
Attrib AGE 0.15437778876366298
Attrib DIS 5.214745586587495
Attrib RAD -1.7212823145503802
Attrib TAX -0.011173609918917245
Attrib PTRATIO 0.47855447929746625
Attrib B -0.8433591800214504
Attrib LSTAT 4.609985528669568
Class
Input
Node 0
Time taken to build model: 8.03 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.8731
Mean absolute error 3.2191
Root mean squared error 4.7344
Relative absolute error 48.2818 %
Root relative squared error 51.3544 %
Total Number of Instances 506
Status

Zastosowania – bardzo dużo …

- Finance
- Marketing
- Economical sciences
- Biology and Medical Science
- Behavioral and social sciences
- Psychology
- Environmental science
- Agriculture
- …

Applications – few words more …

- Finance:
  - The capital asset pricing model uses linear regression as well as the other predictive models for analyzing and quantifying the systematic risk of an investment.
- Marketing
  - Analysis of sales, demands for products, …
- Economical sciences
  - Macro-economical models for countries, etc.
- Biology and Medical Science
  - The scale of illness depeneding on epidemiology indicaters
- Behavioral and social sciences
- Psychology
- Environmental science
  - E.g. to measure the effects of pulp mill or metal mine effluent on the aquatic ecosystem

Literatura

- Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.
- Statystyczne systemy uczące się, Koronacki, Ćwik, WNT 2009 (2 wydanie)
- Statystyka w zarządzaniu, A.Aczel, PWN 2000.
- Metody i modele eksploracji danych. D. Larose, PWN 2008.
- Przystępny kurs statystyki, Stanisz A., 1997 (kol. wyd.)
  - Tom 2 → poświęcony wyłącznie analizie regresji!
- Statystyka. Ekonometria. Prognozowanie. Ćwiczenia z Excelem. A. Snarska, Wydawnictwo Placet 2005.
- I wiele innych …
- Przykłady zastosowań → także czytelnia firmy Statsoft