Predictive modeling - regression

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

an „older” Polish lecture adapted
for Software Engineering Course
Poznan 2009 and updated in 2010

Outline

1. Introduction to regression analysis
    1. Review of basic concepts – linear models
2. Ocena poprawności modelu regresji liniowej.
   (Measuring the quality/fit of the regression model)
3. Regresja wielowymiarowa.
   (Multiple regression)
4. Regresja nieliniowa.
   (Nonlinear regression)
5. Selekcja zmiennych.
   (Feature selection)

- Uwagi: proszę odwołać się do przedmiotu „Statystyka i analiza danych” studia inżynierskie
- Repeat from older notes as above.

Regression analysis – general remarks

- A way of predicting the value of one variable depending on the value of another one (or many others).
  - Dependent vs. independent variables.
- It is a hypothetical model of the relationship between two or more variables.

$$y = f(\mathbf{x}, \beta)$$

- Basic / simple model
  - Linear function $\hat{y} = b_1 \cdot x + b_0$
  - Describe the relationship between variables using the equation of a straight line.
- More sophisticated: non-linear, piecewise family of functions, locally weighted regression, regression trees.

Numeric prediction – Linear regression function

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

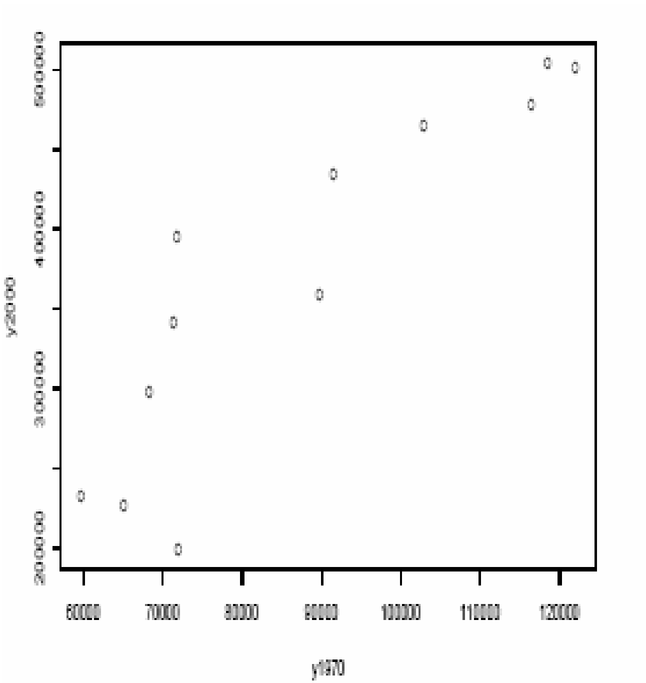
- Linear regression function

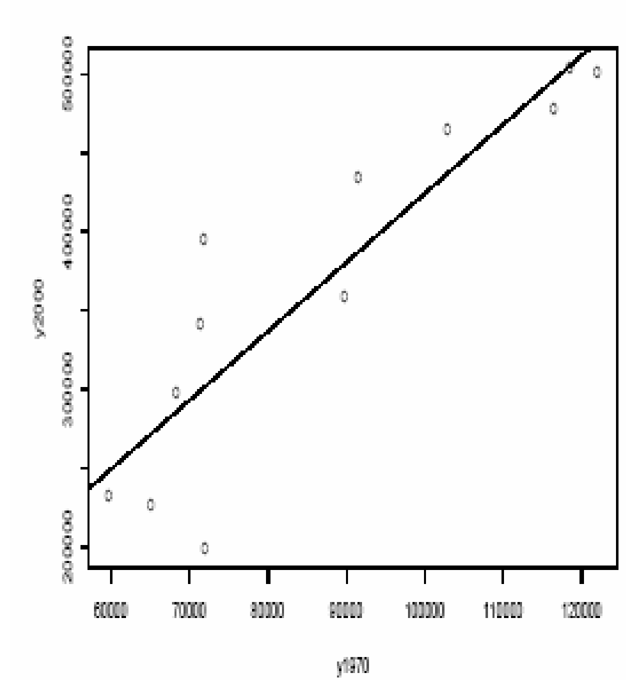
```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Simple example – prices of homes

- Data set *homedata* (from R project) prices of 6841 houses from Maplewood (New Jersey) years: 1970 i 2000.
  Identify relationship between these two variables.

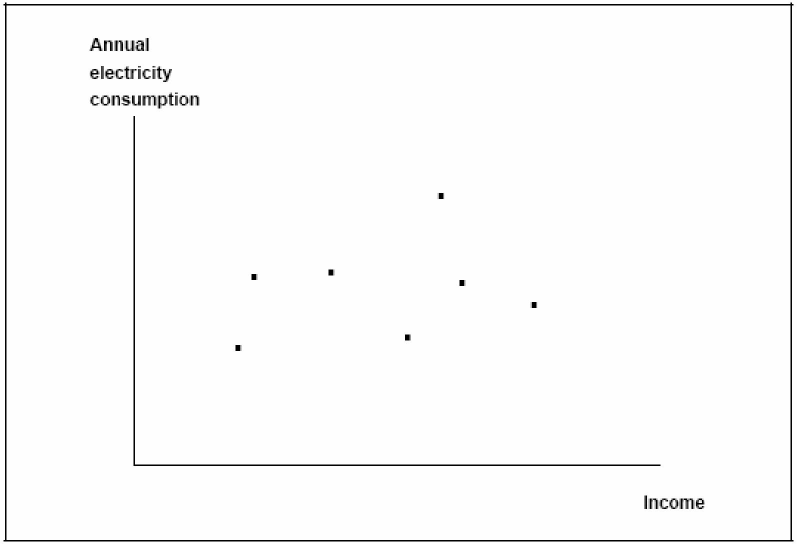
```
> homedata[1:12,]
    y1970  y2000
1    89700 359100
2   118400 504500
3   116400 477300
4   122000 500400
5    91500 433900
6   102800 464800
7    71700 395300
8    71400 340700
9    68200 297400
10   71900 198600
11   65100 225800
12   59700 231500
```

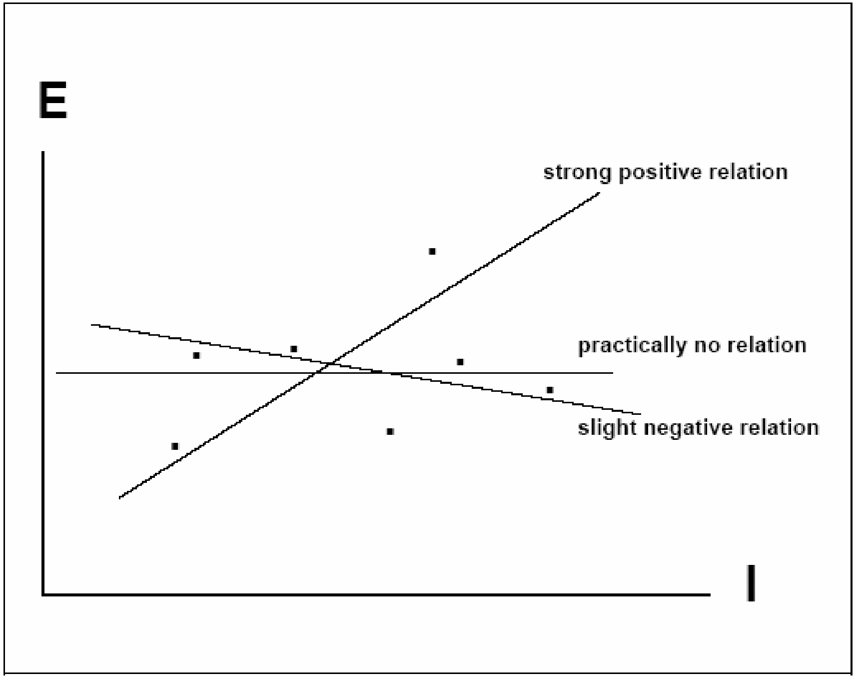




Intuition behind fitting a line

- An example from a lecture on Econometrics (UCI Berkley):
  - Do high income households consume more or less electricity than lower income households?
  - Take a sample of households. Observe the energy consumption and income of each household.
  - Fit the Model that best describes the data!





Which line is the best?

Terminology

- Input variables for prediction – predictors, regressors, predictor variables, independent, explanatory variables
- Output – dependent variable, response or target
- Least squares fitting
- Interpreting the model
- Residuals, R2
- Inference and generalization
- Goodness of the model
- Diagnostics and model inspection

Simple Linear Regression $y = w_0+w_1x$

- <u>Linear regression</u>: involves a response variable y and a single predictor variable x

  $$y = w_0 + w_1 x$$

  where w_0 (y-intercept) and w_1 (slope) are regression coefficients
- <u>Method of least squares</u>: estimates the best-fitting straight line

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

- <u>Multiple linear regression</u>: involves more than one predictor variable
  - Training data is of the form $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_{|D|}}, y_{|D|})$
  - Ex. For 3-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
  - Solvable by extension of least square method
  - Many nonlinear functions can be transformed into the above

Matrix notation

- More general form

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{b}$$

- Solution (MNK)

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \times \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$\mathbf{y}=\begin{bmatrix}100\\100\\200\\250\\350\end{bmatrix}\quad \mathbf{X}=\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\end{bmatrix}\quad \mathbf{X}^T=\begin{bmatrix}1&1&1&1&1\\1&2&3&4&5\end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X}=\begin{bmatrix}5&15\\15&55\end{bmatrix}\quad \det\mathbf{X}^T\mathbf{X}=50$$

$$(\mathbf{X}^T\mathbf{X})^{-1}=\begin{bmatrix}1{,}1&-0{,}3\\-0{,}3&0{,}1\end{bmatrix}\quad \mathbf{X}^T\mathbf{y}=\begin{bmatrix}1000\\3650\end{bmatrix}$$

$$\mathbf{b}=\begin{bmatrix}5\\65\end{bmatrix}$$

$$\hat{\mathbf{y}}=5+65x$$

Interpretation of the linear model

The coefficient in a multiple regression model

- If the j-th predictor variable xj is increased by one unit while all the other predictor variables are kept fixed, then the response variable y will increase by bj.
- A kind of conditional effect!

However – a restrictive assumption on independency between predictors

- In practice – one variable may partly depend on other
- Especially important if one constructs the model in a sequential manner
- The size of a coefficient – approx. of relative importance of variables

How to do it in Excel?
Funkcje stat. REGLINP or dodatek Analiza Danych

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotność R	0,96958969
R kwadrat	0,940104167
Dopasowany R kwa	-1,4
Błąd standardowy	0,579151678
Obserwacje	1

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	7	26,32291667	3,760417	78,47826	0
Resztkowy	5	1,677083333	0,335417		
Razem	12	28			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95,0%*	*Górne 95,0%*
Przecięcie	-0,75	0,579151678	-1,295	0,251891	-2,23875	0,738754	-2,23875	0,738754
Zmienna	1,385416667	0,156388827	8,858796	0,000305	0,983407	1,787426	0,983407	1,787426

współczynnik a

wyraz wolny

Rozkład normalny

☐ Rozkład prawdopodobieństwa normalnego

But – how to interpret these results?

Another examples height = *f*(age) / Statistica (Statsoft)

Dane: Re...

	1 WIEK	2 WZROST
1	7,0	120
2	8,0	122
3	9,0	125
4	10,0	131
5	11,0	135
6	11,5	140
7	12,0	142
8	13,0	145
9	14,0	150
10	15,0	154
11	16,0	159
12	17,0	162
13	18,0	164
14	18,5	168
15	19,0	170

Wartości przewidywane i reszty (regrwzrost15.st

Dalej...

Zmienna zależna: WZROST

Nr przypa	Obserwow Wartość	Przew. Wartość	Reszta
1	120,0000	118,8229	1,17710
2	122,0000	123,1278	-1,12775
3	125,0000	127,4326	-2,43261
4	131,0000	131,7375	-,73747
5	135,0000	136,0423	-1,04233
6	140,0000	138,1947	1,80525
7	142,0000	140,3472	1,65282
8	145,0000	144,6520	,34796
9	150,0000	148,9569	1,04311
10	154,0000	153,2617	,73825
11	159,0000	157,5666	1,43340
12	162,0000	161,8715	,12854
13	164,0000	166,1763	-2,17633
14	168,0000	168,3288	-,32875
15	170,0000	170,4812	-,48119

r. R : ,99684240 F = 2048,784
R^2: ,99369478 df = 1,13
r. R^2: ,99320976 p = ,000000
macji: 1,389446435
l std.: 1,311759 t(13) = 67,611 p < ,0000

V: WZROST (regrwzrost15.sta)

	df	Średnia kwadrat.	F	poziom p
	1	3955,303	2048,784	,000000
	13	1,931		

Podsumowanie regresji — Predykcja zmiennej zal. — OK

Analiza wariancji — Oblicz granice ufności — Anuluj

Oblicz granice predykcji

Kowariancja wsp. regresji — Alfa: ,05 — Analiza reszt

Korelacje i stat. opisowe

Zastosuj

Podsumowanie regresji zmiennej zależnej: WZROST

REGRESJA WIELOKR.

R= ,99684240 R2= ,99369478 Popraw. R^2= ,99320976

F(1,13)=2048,8 p<,00000 Błąd std. estymacji: 1,3894

N=15	BETA	Błąd st. BETA	B	Błąd st. B	t(13)	poziom p
W. wolny			88,68890	1,311759	67,61067	,000000
WIEK	,996842	,022023	4,30486	,095107	45,26349	,000000

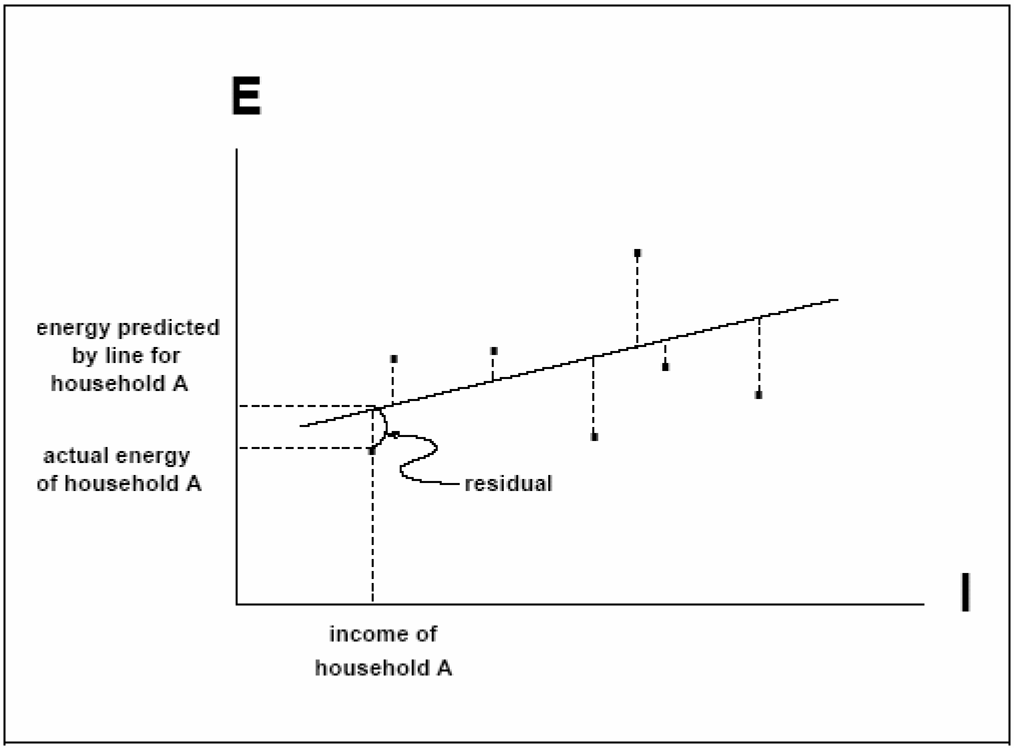
How Good is the Model?

- The regression function is only a model based on the data.
- This model might not reflect reality (in particular in case of a sample + imperfect data points)
  - We need some way of testing how well the model fits the observed data.

  - How? What kind of tools?

Residual

- Residuals $e_i = y_i - \hat{y}_i$
- For a good fit we need residual to be small



Basic concepts of verifying linear model

- Study the fittnes of the calcuted linear model to the changing points x,y.
- Składnik resztowy (residuals) $e_i = y_i - \hat{y}_i$

  tym większy, im większy jest składnik losowy ε,

  może także wynikać z błędnego przyjęcia danej funkcji regresji.

Analyse całkowitej zmienności y – variance analysis

- Oceniamy za pomocą wariancji S_y^2 lub całkowitej sumy kwadratów różnic SST

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

  - Variance of y is a function of the mean value of y, not a constant

Basic formula used in other tools

- Całkowitą sumę kwadratów odchyleń (*SST*) w analizie regresji dzieli się na dwie części:

$$SST = SSR + SSE$$

- gdzie $\sum(y-\overline{y})^2 = \sum(\hat{y}-\overline{y})^2 + \sum(y-\hat{y})^2$
- *SSR* – regresyjna suma kwadratów odchyleń (część wyjaśniona przez zbudowany model),
- *SSE* – resztowa suma kwadratów odchyleń (część nie wyjaśniona przez zbudowany model).

- SS_T
  - Total variability (variability between scores and the mean).
- SS_R
  - Residual/Error variability (variability between the regression model and the actual data).
- SS_M
  - Model variability (difference in variability between the model and the mean).

R2 – współczynnik determinacji

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

- Najważniejsza miara dopasowania funkcji regresji do danych empirycznych;
- Współczynnik determinacji (determination coefficient) --- przyjmuje wartości z przedziału [0,1] i wskazuje jaka część zmienności zmiennej *y* jest wyjaśniana przez znaleziony model. Na przykład dla R^2=0.619 znaleziony model wyjaśnia około 62% zmienności *y*.
- Range [0,1]; the higer, the better
- R^2 gives percentage of variation in dependent variable that is explained by the model

Przy okazji: pomyśl o związku współczynnika R^2 oraz współczynnika korelacji *r*.

Other related measures

- Współczynnik determinacji:

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

- Jest to stosunek zmienności wyjaśnianej przez model do zmienności całkowitej.
- Medium Square Error -Średni błąd kwadratowy:

$$MSE = \frac{SSE}{n-2}$$

- Wariancja resztowa (k liczba zmiennych)

$$S_e^2 = \frac{1}{n-(k+1)} \sum_i e_i^2$$

- (Standard errors) Błędy standardowe parametrów *bi* :

$$S(b_j) = \sqrt{S_e^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1}} = S_e \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$

$$S(b_1) = \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$S(b_0) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- Standard deviation - odchylenie standardowe składnika resztowego – standardowy błąd oszacowania

$$S = \sqrt{\frac{SSE}{n-2}}$$

Adjusted R2 – skorygowany współczynnik determinacji

- This is the coefficient of determination adjusted for the degrees of freedom.
- It has been adjusted to take into account the sample size and the number of independent variables. If the number of independent variables, *k*, is large relative to the sample size *n*, the unadjusted R2 may be unrealistically high!

Inference and generalization

- Sample vs. the rest of population
  - Predictions for objects x we do not know their y values.
- Goodness of fit to the training data is not an objective.
- Coefficient in our model may capture peculiarities of the training sample
- Basic tool – test hypotheses whether the population regression coefficients!
  - Global level vs. local …

Testing hypothesis on coefficients

- We need to test whether the population regression coefficients β are really zero!
- The observed data produces a model by chance, even there was no structure (variables were not related) in the population the data were collected from!
- Global test (at least one coefficient is significant)

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m$$

$$H_1 : \exists i : \beta_i \neq 0$$

$$F = \frac{SSR/m}{SSE(n-m-1)}$$

Local test – a question about particular coefficient

- Evaluating single parameters β_i in regression model (if *y* is linearly dependent on *x*) – local test.
- Hypotheses

$$H_0: \quad \beta_i = 0$$

$$H_1: \quad \beta_i \neq 0$$

- Test statistics:

$$t = \frac{\beta_i}{S(\beta_i)}$$

Again example height = *f*(age) / Statistica (Statsoft)

Dane: Re...

	1 WIEK	2 WZROST
1	7,0	120
2	8,0	122
3	9,0	125
4	10,0	131
5	11,0	135
6	11,5	140
7	12,0	142
8	13,0	145
9	14,0	150
10	15,0	154
11	16,0	159
12	17,0	162
13	18,0	164
14	18,5	168
15	19,0	170

Wartości przewidywane i reszty (regrwzrost15.st

Dalej...

Zmienna zależna: WZROST

Nr przypa	Obserwow Wartość	Przew. Wartość	Reszta
1	120,0000	118,8229	1,17710
2	122,0000	123,1278	-1,12775
3	125,0000	127,4326	-2,43261
4	131,0000	131,7375	-,73747
5	135,0000	136,0423	-1,04233
6	140,0000	138,1947	1,80525
7	142,0000	140,3472	1,65282
8	145,0000	144,6520	,34796
9	150,0000	148,9569	1,04311
10	154,0000	153,2617	,73825
11	159,0000	157,5666	1,43340
12	162,0000	161,8715	,12854
13	164,0000	166,1763	-2,17633
14	168,0000	168,3288	-,32875
15	170,0000	170,4812	-,48119

r. R : ,99684240 F = 2048,784
R^2: ,99369478 df = 1,13
r. R^2: ,99320976 p = ,000000
macji: 1,389446435
l std.: 1,311759 t(13) = 67,611 p < ,0000

V: WZROST (regrwzrost15.sta)

	df	Średnia kwadrat.	F	poziom p
	1	3955,303	2048,784	,000000
	13	1,931		

Podsumowanie regresji

Predykcja zmiennej zal.

OK

Oblicz granice ufności

Oblicz granice predykcji

Anuluj

Analiza wariancji

Alfa: ,05

Kowariancja wsp. regresji

Analiza reszt

Zastosuj

Podsumowanie regresji zmiennej zależnej: WZROST

REGRESJA WIELOKR.

R= ,99684240 R2= ,99369478 Popraw. R^2= ,99320976

F(1,13)=2048,8 p<,00000 Błąd std. estymacji: 1,3894

N=15	BETA	Błąd st. BETA	B	Błąd st. B	t(13)	poziom p
W. wolny			88,68890	1,311759	67,61067	,000000
WIEK	,996842	,022023	4,30486	,095107	45,26349	,000000

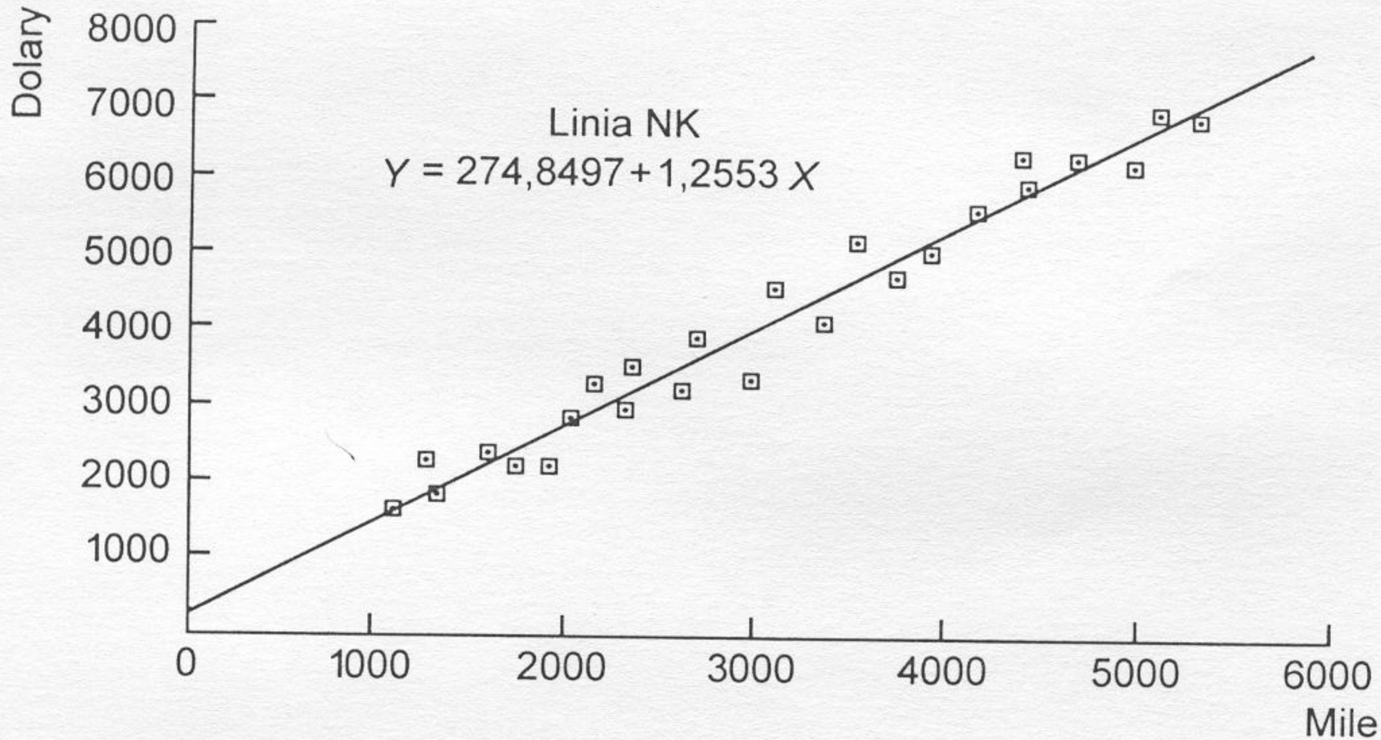
American Express - case

- Rozważmy przykład posiadaczy kart kredytowych American Express → firma jest przekonana, że posiadacze jej kart podróżują więcej niż inni ludzie.
- W badaniach marketingowych podjęto próbę ustalenie związków między długością tras podróży a obciążeniem karty kredytowej jej posiadacza w danym okresie czasu.
- Więcej w Aczel: Statystyka w zarządzaniu, str. 468.

Tablica 10.1. Dane do badania przeprowadzonego na zlecenie American Express

Długość tras (w milach)	Obciążenie kart (w $)
1 211	1 802
1 345	2 405
1 422	2 005
1 687	2 511
1 849	2 332
2 026	2 305
2 133	3 016
2 253	3 385
2 400	3 090
2 468	3 694
2 699	3 371
2 806	3 998
3 082	3 555
3 209	4 692
3 466	4 244
3 643	5 298
3 852	4 801
4 033	5 147
4 267	5 738
4 498	6 420
4 533	6 059
4 804	6 426
5 090	6 321
5 233	7 026
5 439	6 964

Calculating a linear regression – American Express



Rysunek 10.12. Linia NK w badaniu zleconym przez American Express

Model diagnostics

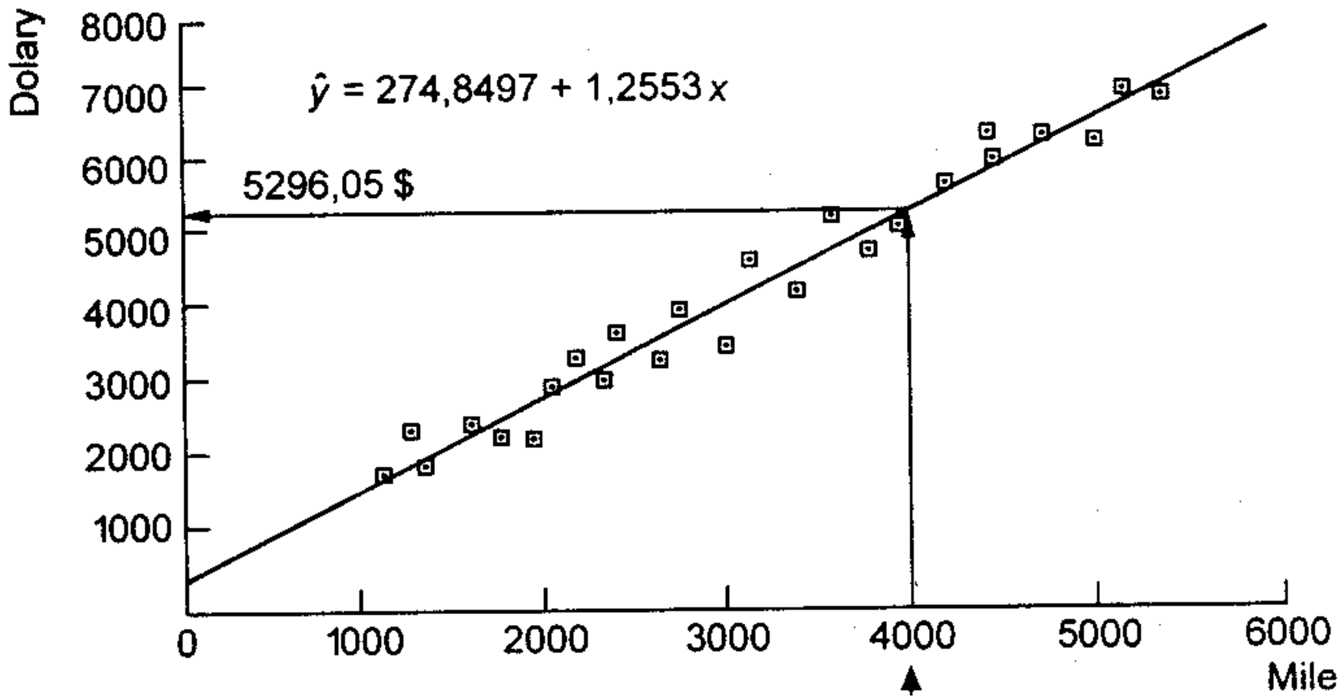
- SSE=2328161,2 MME=SSE/(n-2) = 101224,4
- Standard error $s = \sqrt{MSE} = 318{,}158$
- Błędy estymacji $S(b_0)$ = 170,338

  $S(b_1)$ = 0.00497
- Współczynnik determinacji R^2 = 0.9652
- All coefficient are statistically significant (globally and locally) for α = 0.05

Predicition with regression model

- Introduce a new point x into the formula.
- What are the expenses of travelers with 4000 miles

$$\hat{y} = 274{,}85 + 1{,}2663 \cdot x = 274{,}85 + 1{,}2663 \cdot 4000 = 5296{,}05$$



Intervals for predictors

- (1-α)$\cdot$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:

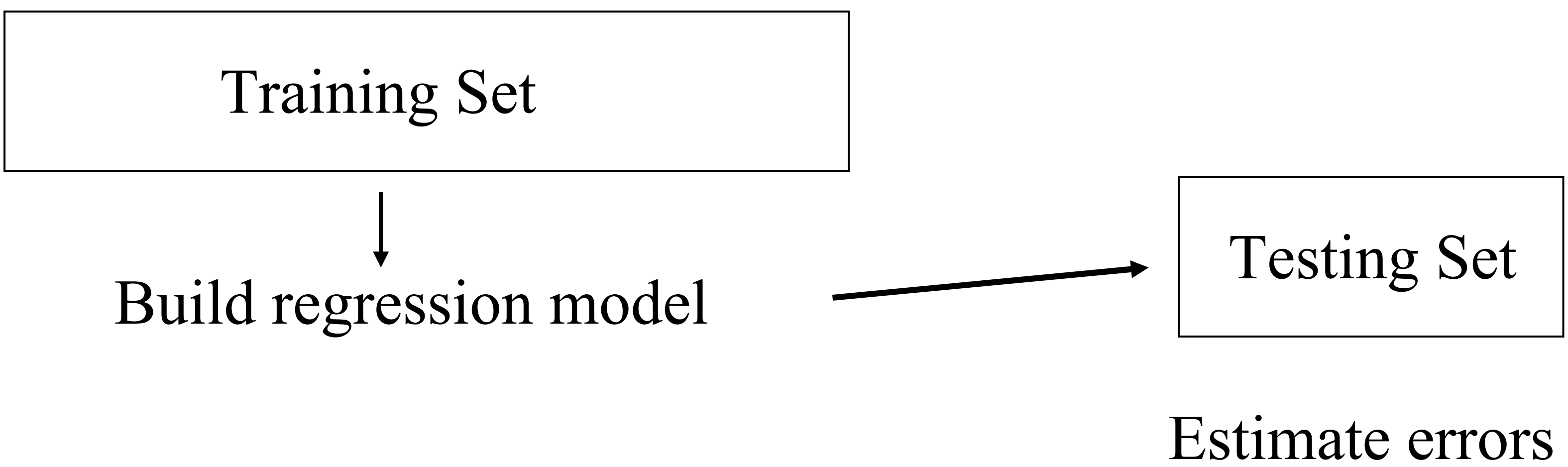
  $\bar{x}$ = 79448/25=3177,92; SS_x = 40947557,84 a *s* = 318,16
  Ponadto *t* przy 23 stopniach swobody wynosi 2,069

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

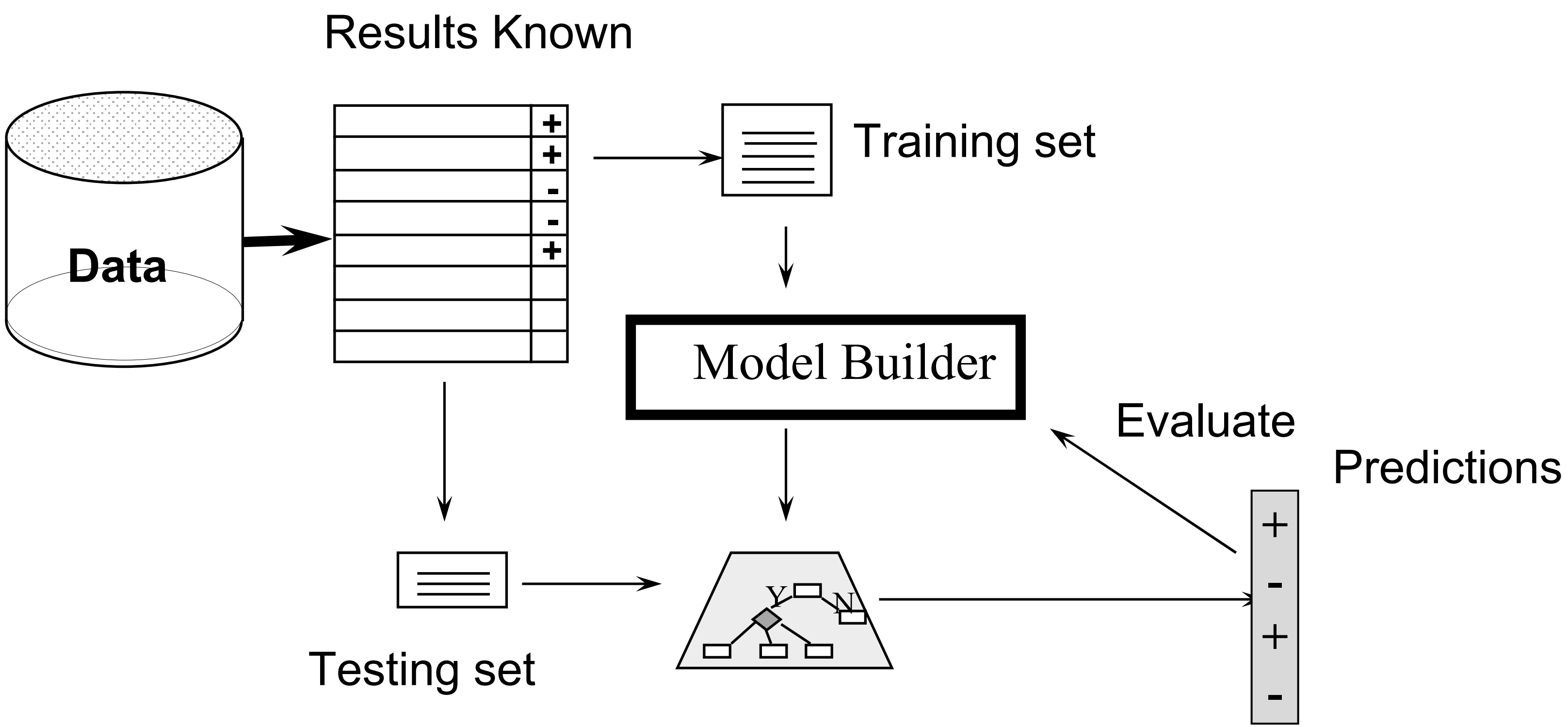
- Oznacza to, że w oparciu o wyniki badań można mieć 95% zaufania do prognozy, że posiadacz karty, który przebył trasę 4000 mil w okresie o danej długości obciąży swoją kartę kredytową sumą od 4619.43 do 5972,67$.

Another view on prediction of testing examples

- Focus on new coming observation – just predication
- Either given set of new / testing data or
- Empirical split (random)



Similar procedure for evaluating classifiers



W odróżnieniu od klasyfikacji nie ma etykiet dyskretnych klas, lecz wynik liczbowy zmiennej zależnej

Other Error Measures

- Measure predictor accuracy: measure how far off the predicted value is from the actual known value
- **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
  - 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

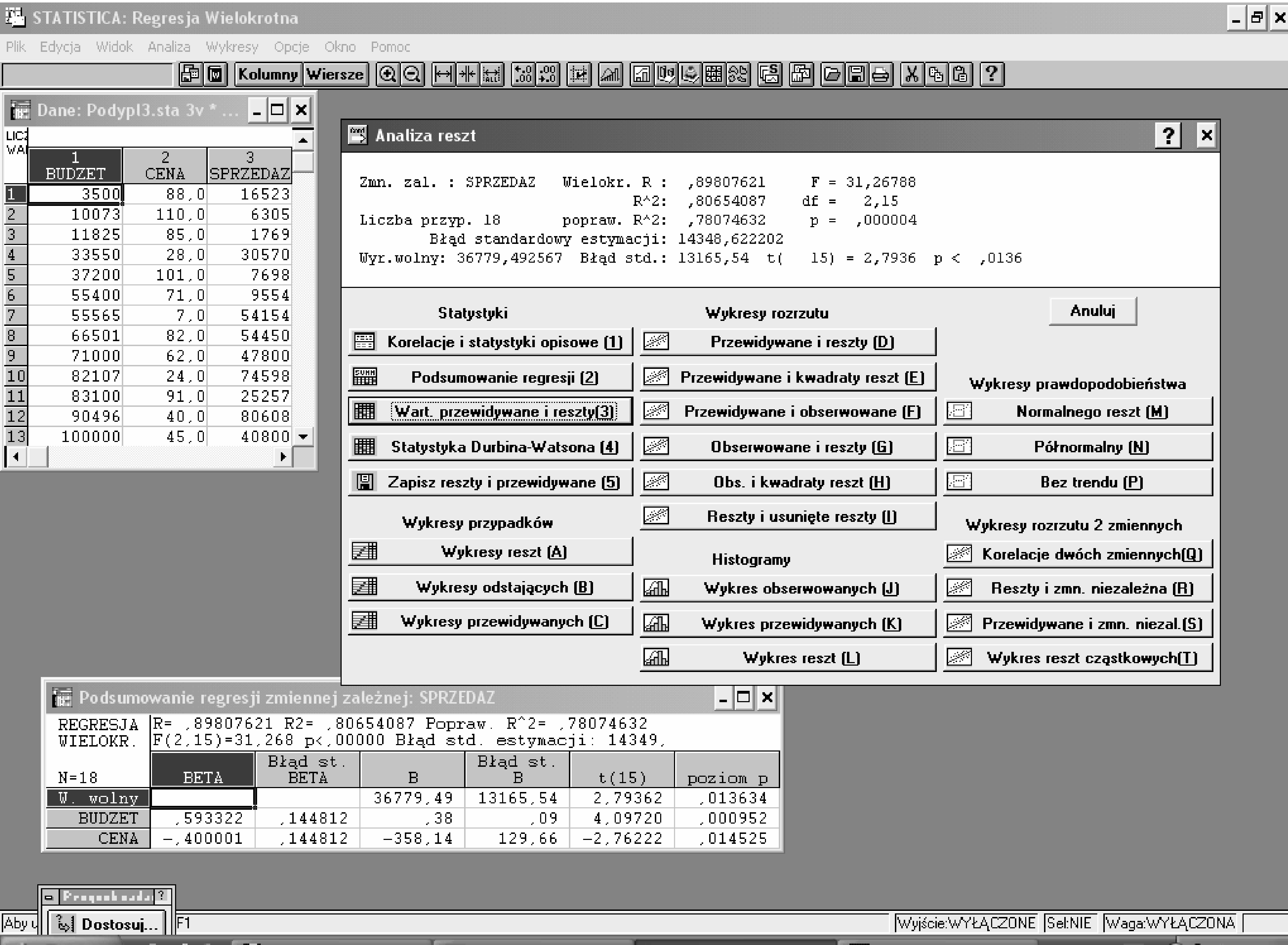
What Conditions Must Hold in order for us to Legitimately Apply Regression Techniques?

- The error variable in the model, ε, must be normally distributed.
- The mean of the error variable must be zero.
- The standard deviation of the error variable must be constant.
- The errors are independent.

- Ciekawa dyskusja założeń w A.Aczel „Statystyka w zarządzaniu”.

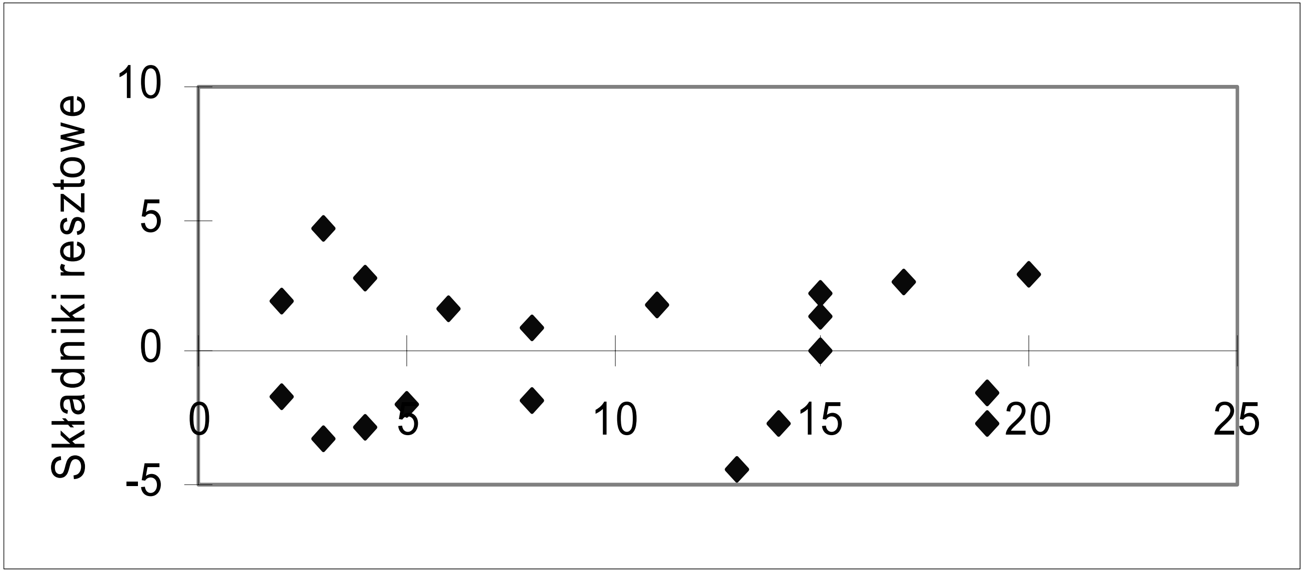
Diagnostics of residual plots

- Simple tool for graphical inspection of the linear regression model
- Checking assumptions:
  - The mean of the error variable (residual) must be zero.
  - The standard deviation of the error variable must be constant.
  - The figure $\varepsilon = f(\hat{y})$ should resemble a characteristic pattern
- The error variable in the model, ε, must be normally distributed – quintiles probability plot


STATISTICA: Regresja Wielokrotna
Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc
Kolumny Wiersze
Dane: Podypl3.sta 3v * ...
1 BUDZET
2 CENA
3 SPRZEDAZ
1 3500 88,0 16523
2 10073 110,0 6305
3 11825 85,0 1769
4 33550 28,0 30570
5 37200 101,0 7698
6 55400 71,0 9554
7 55565 7,0 54154
8 66501 82,0 54450
9 71000 62,0 47800
10 82107 24,0 74598
11 83100 91,0 25257
12 90496 40,0 80608
13 100000 45,0 40800
Analiza reszt
Zmn. zal. : SPRZEDAZ Wielokr. R : ,89807621 F = 31,26788
R^2: ,80654087 df = 2,15
Liczba przyp. 18 popraw. R^2: ,78074632 p = ,000004
Błąd standardowy estymacji: 14348,622202
Wyr.wolny: 36779,492567 Błąd std.: 13165,54 t(15) = 2,7936 p < ,0136
Statystyki
Korelacje i statystyki opisowe (1)
Podsumowanie regresji (2)
Wart. przewidywane i reszty(3)
Statystyka Durbina-Watsona (4)
Zapisz reszty i przewidywane (5)
Wykresy przypadków
Wykresy reszt (A)
Wykresy odstających (B)
Wykresy przewidywanych (C)
Wykresy rozrzutu
Przewidywane i reszty (D)
Przewidywane i kwadraty reszt (E)
Przewidywane i obserwowane (F)
Obserwowane i reszty (G)
Obs. i kwadraty reszt (H)
Reszty i usunięte reszty (I)
Histogramy
Wykres obserwowanych (J)
Wykres przewidywanych (K)
Wykres reszt (L)
Anuluj
Wykresy prawdopodobieństwa
Normalnego reszt (M)
Półnormalny (N)
Bez trendu (P)
Wykresy rozrzutu 2 zmiennych
Korelacje dwóch zmiennych(Q)
Reszty i zmn. niezależna (R)
Przewidywane i zmn. niezal.(S)
Wykres reszt cząstkowych(T)
Podsumowanie regresji zmiennej zależnej: SPRZEDAZ
REGRESJA WIELOKR.
R= ,89807621 R2= ,80654087 Popraw. R^2= ,78074632
F(2,15)=31,268 p<,00000 Błąd std. estymacji: 14349,
N=18 BETA Błąd st. BETA B Błąd st. B t(15) poziom p
W. wolny 36779,49 13165,54 2,79362 ,013634
BUDZET ,593322 ,144812 ,38 ,09 4,09720 ,000952
CENA -,400001 ,144812 -358,14 129,66 -2,76222 ,014525
Dostosuj...
F1
Wyjście:WYŁĄCZONE Sel:NIE Waga:WYŁĄCZONA

Residual graphs

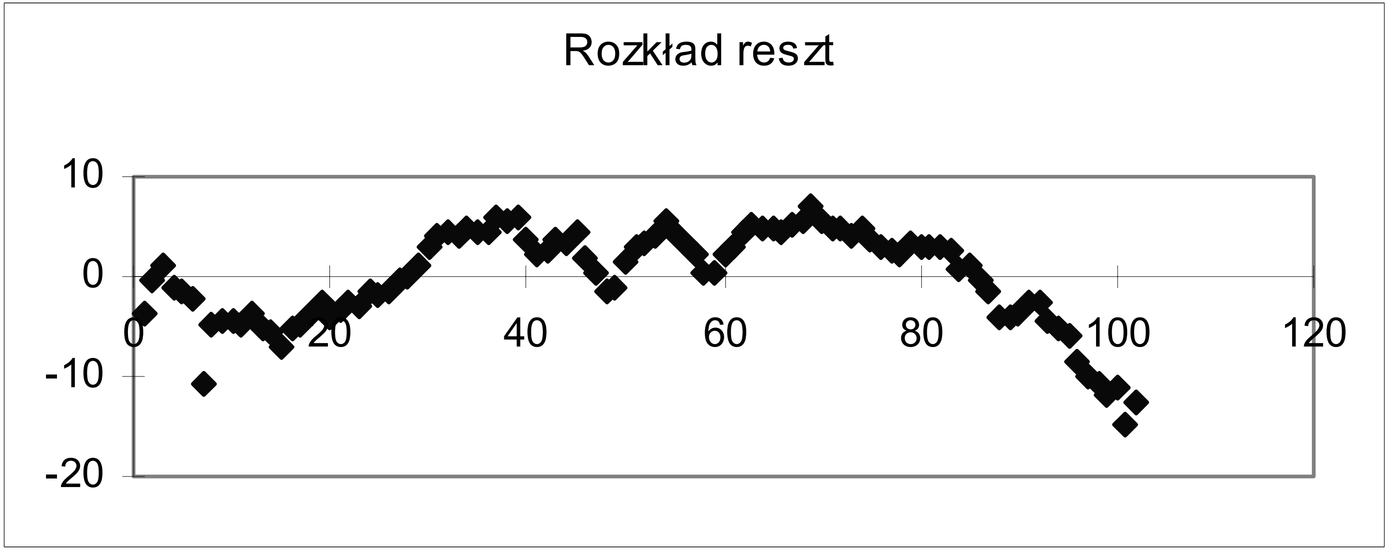
- The figure $\varepsilon=f(\hat{y})$



- Standarized residuals should be located in a kind of „belt" shape – approx. equally distributed around the expected 0.

Other shape of residual graph

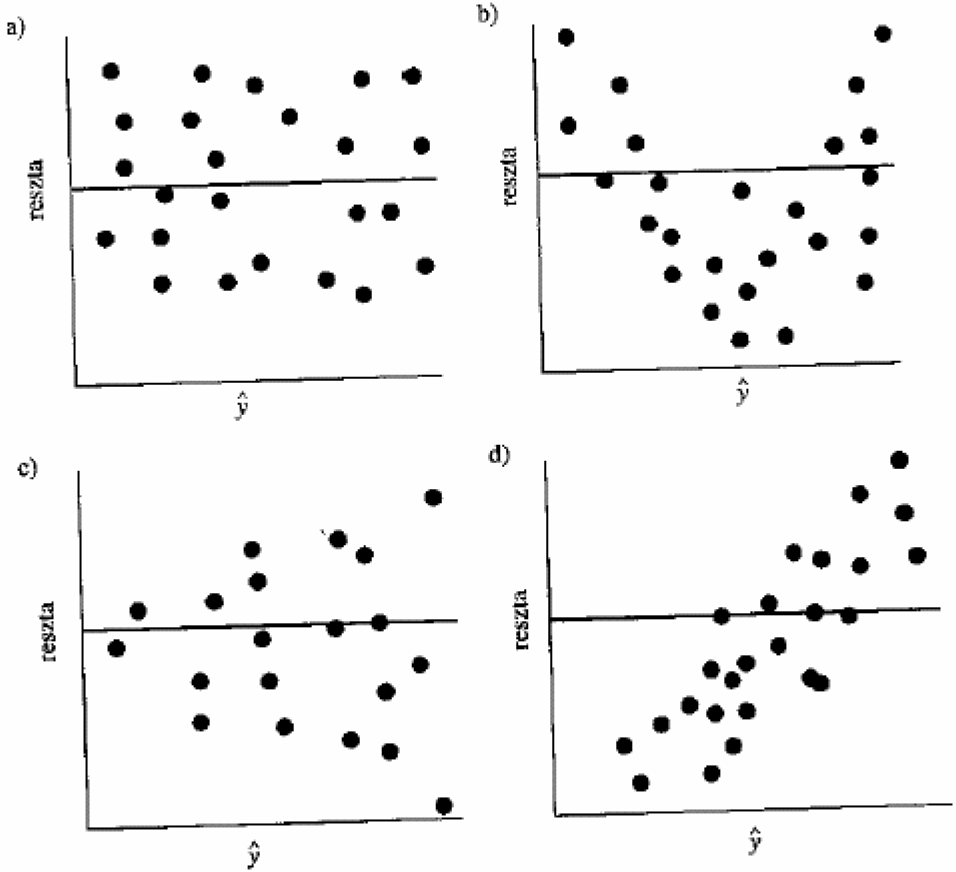
- Consider another examples.



- Residuals are not located inside the regular shape + we can suspect that some points are strongly auto-corelated (test Durbina-Watsona).

Interpret other figures

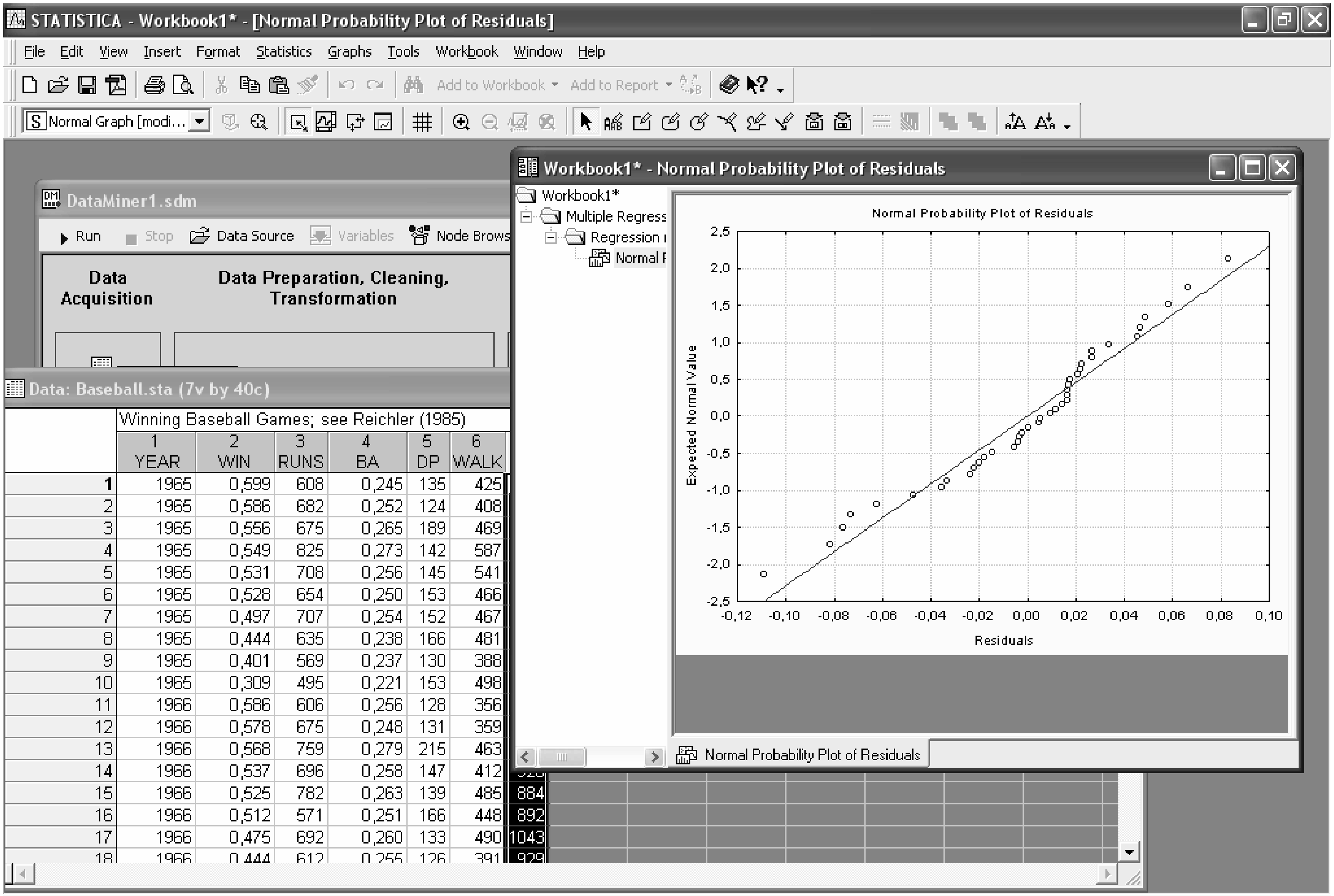
- Different situations



Rys. 2.13. Cztery możliwe układy punktów na wykresach reszt względem wartości przewidywanych

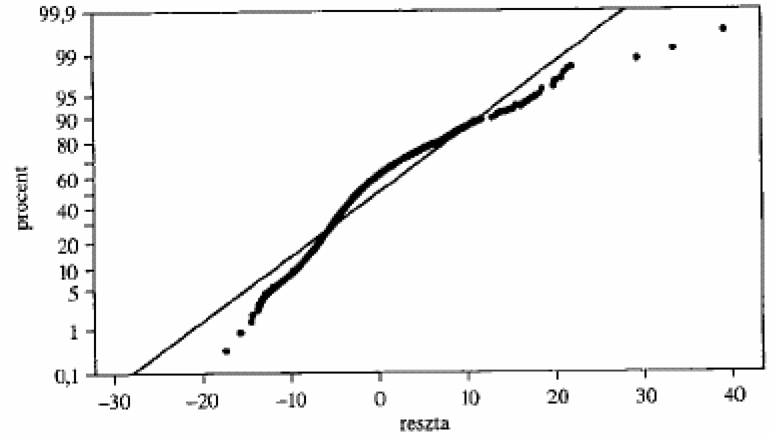
Inspecting normality of residul distribution

- Dataminer 7 (Normality Probability Plot of Residuals)

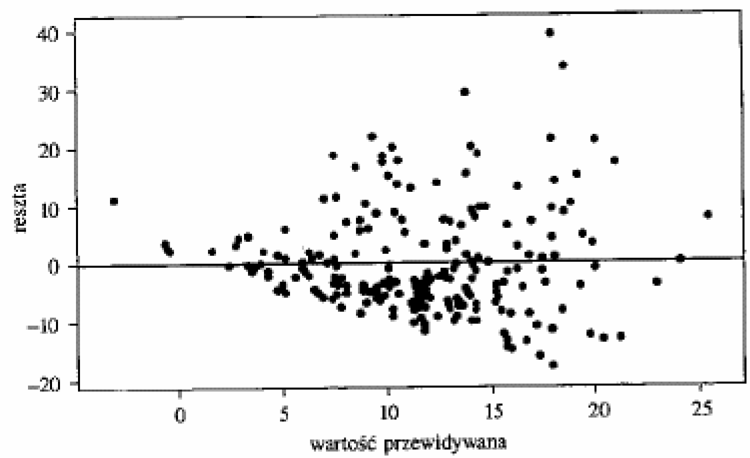


Another examples „baseball American League 2002”

Zależność między średnią uderzeń gracza a liczba uderzeń, które pozwoliły na zaliczenie baz i zdobycie punktu.
[larose 08, § 2.10



Rys. 2.15. Wykres kwantylowy standaryzowanych reszt — naruszone założenie o rozkładzie normalnym

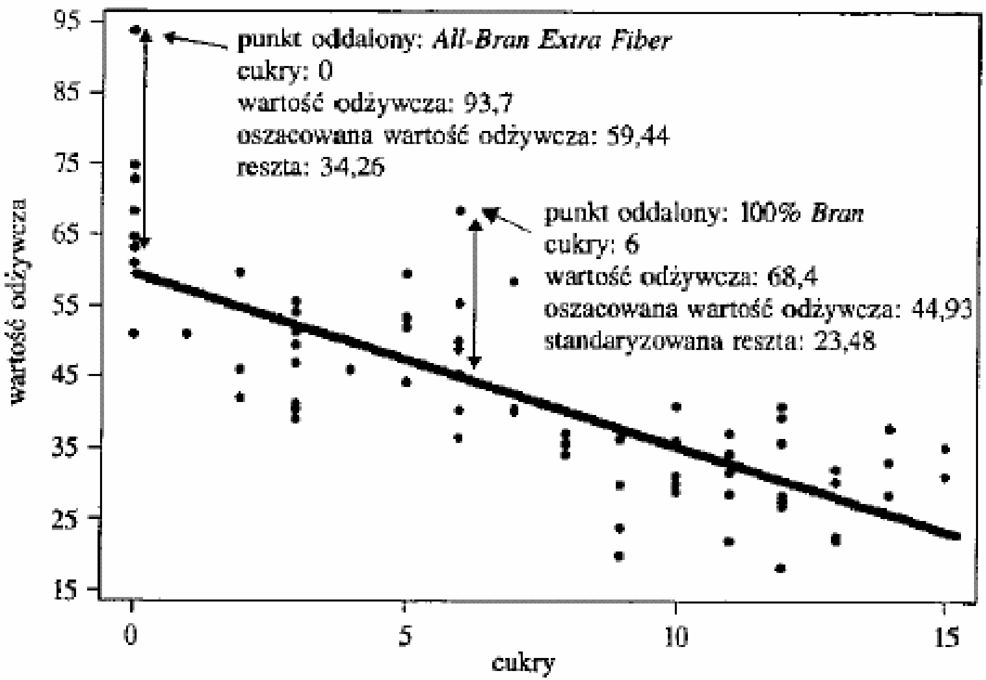


Rys. 2.16. Wykres standaryzowanych reszt względem wartości przewidywanych — naruszone założenie o stałej wariancji

- Both assumptions are violated!

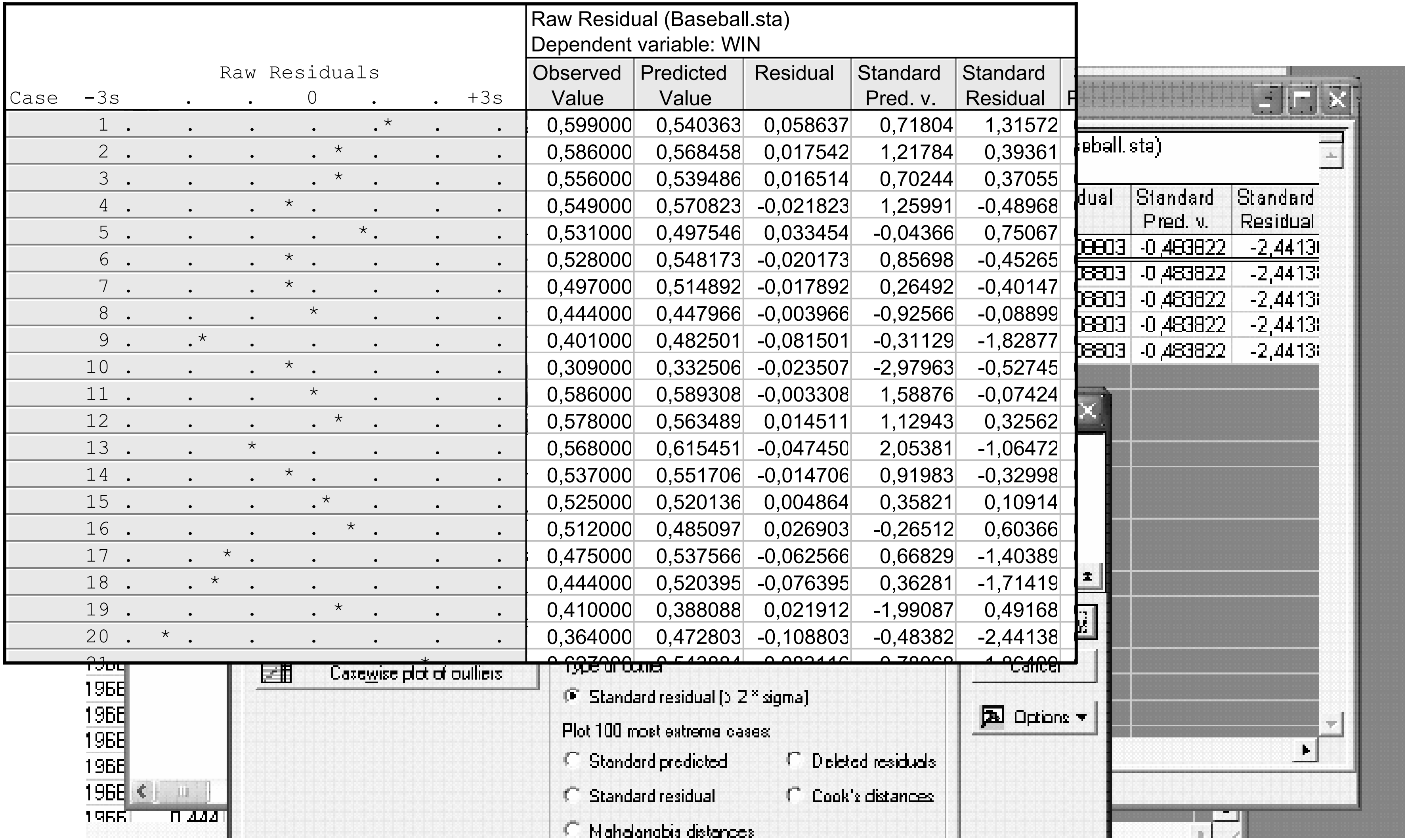
Punkty oddalone - outliers

- Case „płatki śniadaniowe” [Larose 08] – two products are too far from the regression line → outliers (obserwacje oddalone, odstające, samotnicze)



Rys. 2.3. Identyfikacja punktów oddalonych dla regresji zmiennej *wartość odżywcza* względem zmiennej *cukry*

Outliers – standarized residuals



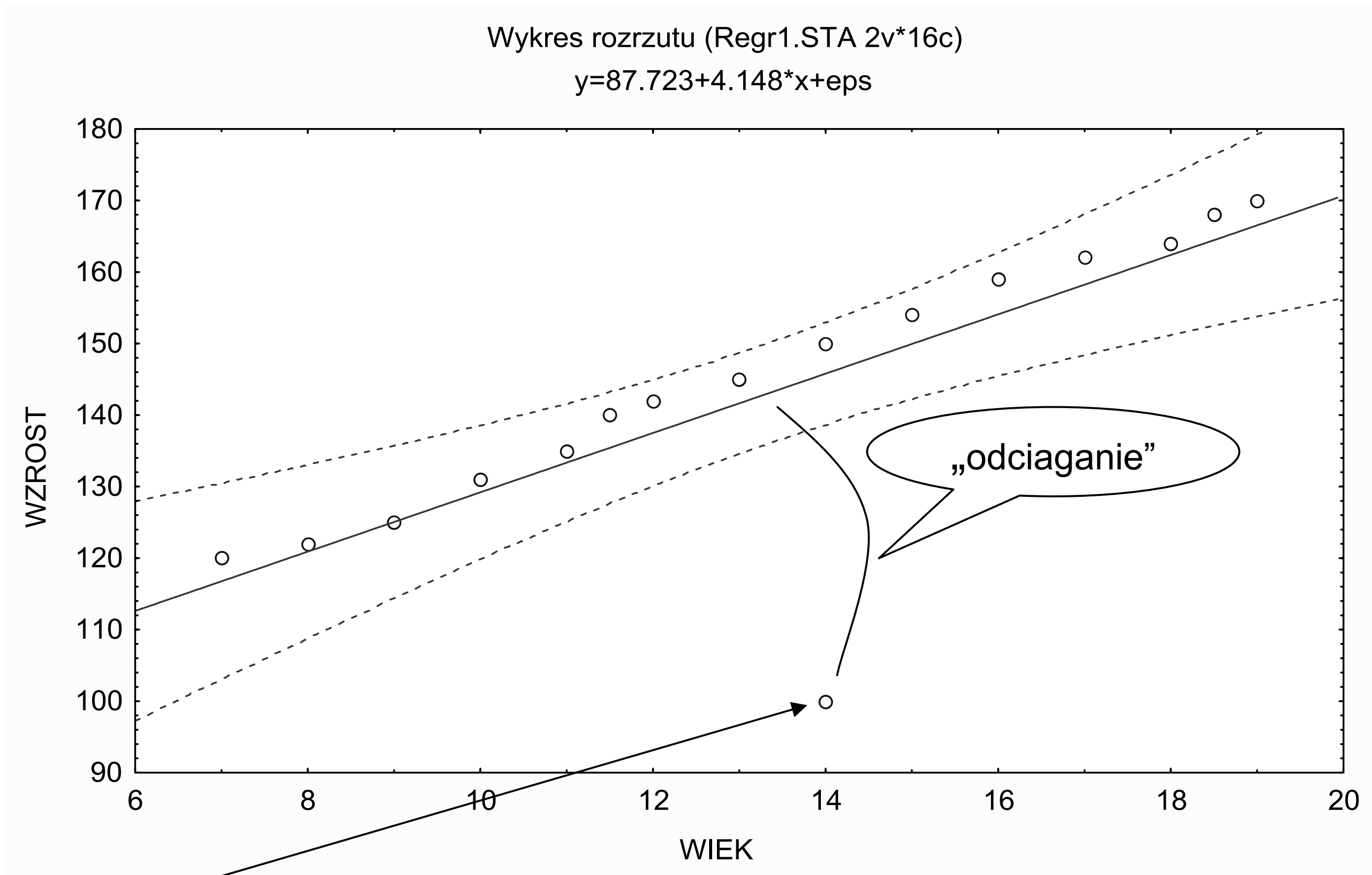
Raw Residual (Baseball.sta)
Dependent variable: WIN

Case	Raw Residuals (-3s . . 0 . . +3s)	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual
1	. . . . .* . .	0,599000	0,540363	0,058637	0,71804	1,31572
2	. . . . * . . .	0,586000	0,568458	0,017542	1,21784	0,39361
3	. . . . * . . .	0,556000	0,539486	0,016514	0,70244	0,37055
4	. . . *	0,549000	0,570823	-0,021823	1,25991	-0,48968
5	. . . . *. . .	0,531000	0,497546	0,033454	-0,04366	0,75067
6	. . . *	0,528000	0,548173	-0,020173	0,85698	-0,45265
7	. . . *	0,497000	0,514892	-0,017892	0,26492	-0,40147
8	. . . * . . .	0,444000	0,447966	-0,003966	-0,92566	-0,08899
9	. .*	0,401000	0,482501	-0,081501	-0,31129	-1,82877
10	. . . *	0,309000	0,332506	-0,023507	-2,97963	-0,52745
11	. . . * . . .	0,586000	0,589308	-0,003308	1,58876	-0,07424
12	. . . . * . . .	0,578000	0,563489	0,014511	1,12943	0,32562
13	. . *	0,568000	0,615451	-0,047450	2,05381	-1,06472
14	. . . *	0,537000	0,551706	-0,014706	0,91983	-0,32998
15	. . . .* . . .	0,525000	0,520136	0,004864	0,35821	0,10914
16	. . . . * . . .	0,512000	0,485097	0,026903	-0,26512	0,60366
17	. . *	0,475000	0,537566	-0,062566	0,66829	-1,40389
18	. . *	0,444000	0,520395	-0,076395	0,36281	-1,71419
19	. . . . * . . .	0,410000	0,388088	0,021912	-1,99087	0,49168
20	. *	0,364000	0,472803	-0,108803	-0,48382	-2,44138

Regresja – the role of outliers

Dane: Re...

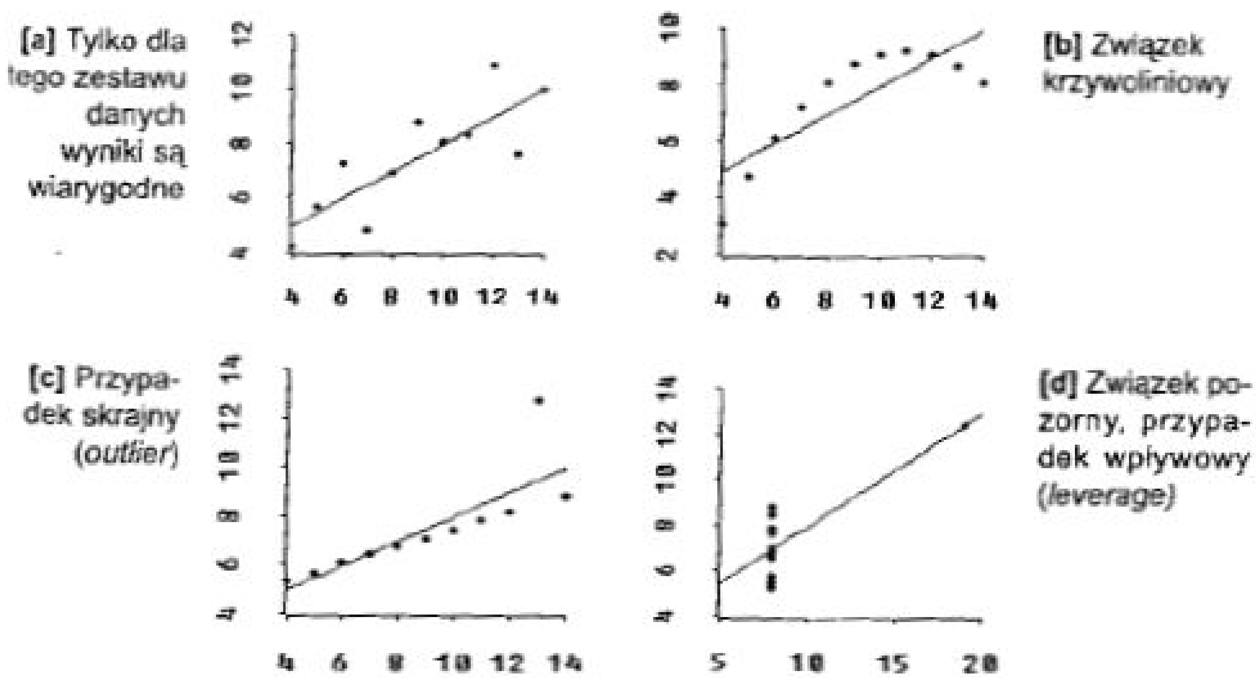
	1 WIEK	2 WZROST
1	7,0	120
2	8,0	122
3	9,0	125
4	10,0	131
5	11,0	135
6	11,5	140
7	12,0	142
8	13,0	145
9	14,0	150
10	14,0	100
11	15,0	154
12	16,0	159
13	17,0	162
14	18,0	164
15	18,5	168
16	19,0	170



Outlier

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 $M_X = 9 \quad M_Y = 7{,}5$, równanie regresji jest dokładnie takie samo. $Y' = 3 + 0{,}5 \times X$.



Rysunek 8.6. Przykład danych Anscombe'a.

Mutliple regression
Regresja wielokrotna (wielowymiarowa, wieloraka)

- Response variable depends on many predictor – quite often in practice and data mining.
- Linear regression model of *y* with respect to *m-1* independent variables $x_1, x_2, \ldots, x_{m-1}$ defined as the following form:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_{m-1} \cdot x_{m-1}$$

- Multidimensional data

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{i1} & x_{i2} & \ldots & x_{im} \end{bmatrix}^T$$

Matrix formulation and Least Square Fitting

- Założenie: wpływ każdej rozpatrywanej zmiennej objaśniającej na zmienną *y* jest liniowy i nie zależy od wartości innych zmiennych

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_{m-1} \cdot x_{m-1} + \varepsilon$$

- Zapis macierzowy: x_m odpowiada *y*; wyraz wolny dodatkowa zmienna $x_{i0} = 1$

$$\underline{Y} = \underline{X} \cdot \underline{\beta} + \underline{\varepsilon}$$

- Rozwiązanie MNK (LSF)

$$\underline{b} = \left(\underline{X}' \cdot \underline{X}\right)^{-1} \cdot \underline{X}' \cdot \underline{Y}$$

Toy examples

Given measurements

x1	5	3	5	3
x2	0,5	0,5	0,3	0,3
f(x1,x2)	1,5	3,5	6,2	3,2

Assume linear regression model

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

Calculations

$$\mathbf{y} = \begin{bmatrix} 1{,}5 \\ 3{,}5 \\ 6{,}2 \\ 3{,}2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 5 & 0{,}5 \\ 1 & 3 & 0{,}5 \\ 1 & 5 & 0{,}3 \\ 1 & 3 & 0{,}3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{bmatrix}$$

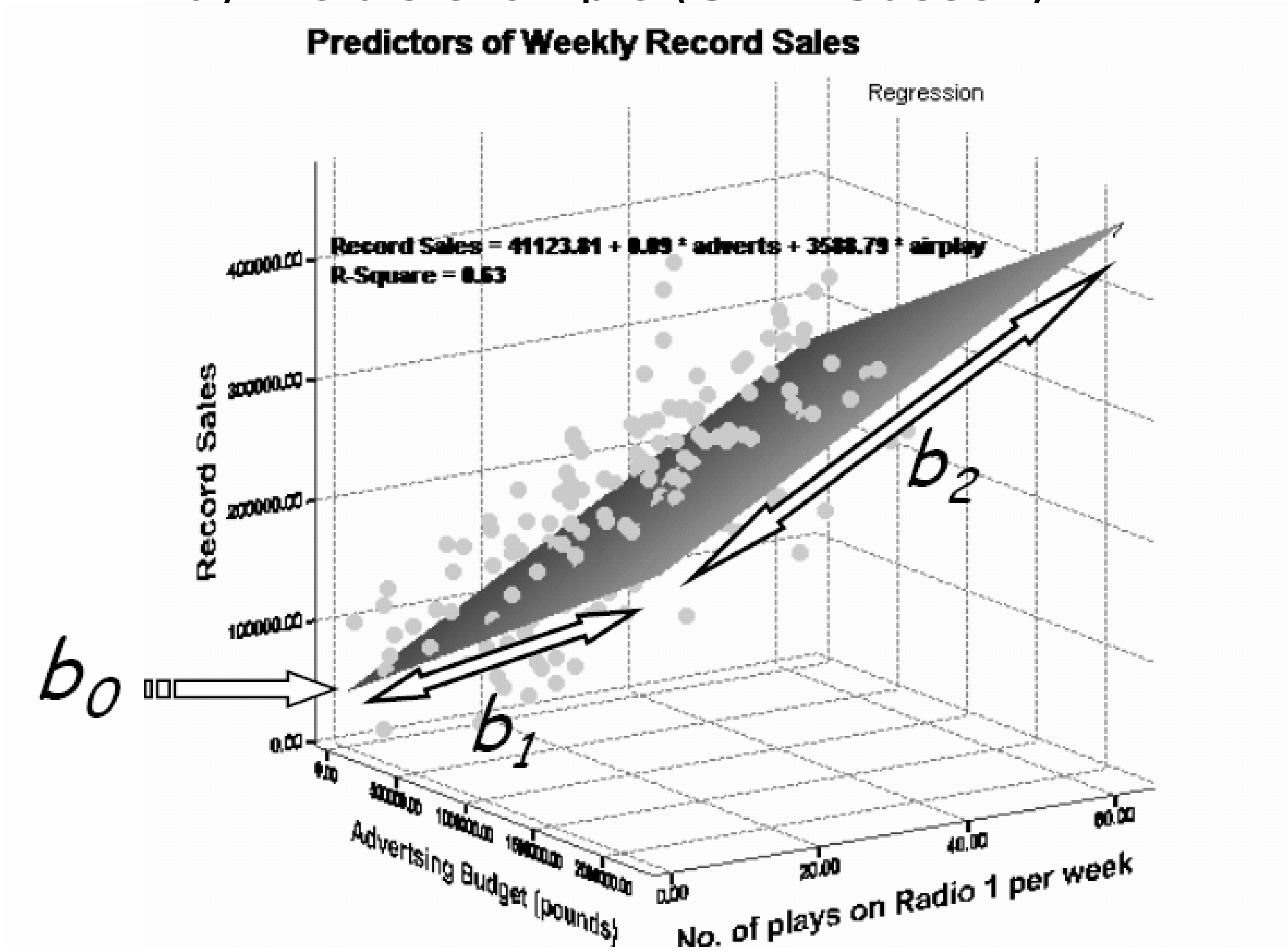
- So finally:

$$\begin{aligned} b_0 &= 7{,}0 \\ b_1 &= 0{,}25 \\ b_2 &= -11 \end{aligned}$$

- Remarks: number of observations *n* should be greater (at least ?) than parameters to be estimated *m*.
- Columuns of matrix – linear independence → unique solution
- Diagnostics → as previous methododology

An example of multiple regression

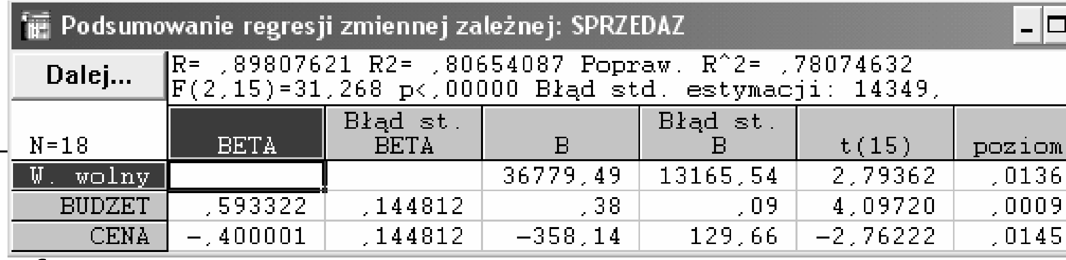
- Andy Field's example (Univ. Sussex)



Statistica Example

- Final sale of the product (sprzedaz)
- Advertisement budget and a price of single item

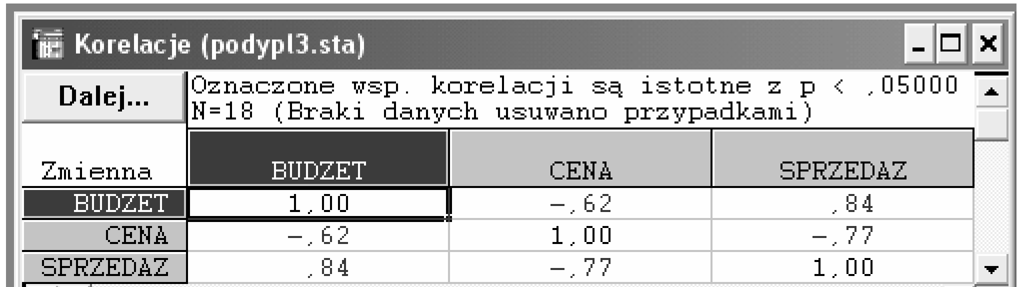
	BUDŻET	CENA	SPRZEDAZ
1	3500	88	16523
2	10073	110	6305
3	11825	85	1769
4	33550	28	30570
5	37200	101	7698
6	55400	71	9554
7	55565	7	54154
8	66501	82	54450
9	71000	62	47800
10	82107	24	74598
11	83100	91	25257
12	90496	40	80608
13	100000	45	40800
14	102100	21	63200
15	132222	40	69675
16	136297	8	98715
17	139114	63	75886
18	165575	5	83360


Podsumowanie regresji zmiennej zależnej: SPRZEDAZ

Dalej...

R= ,89807621 R2= ,80654087 Popraw. R^2= ,78074632
F(2,15)=31,268 p<,00000 Błąd std. estymacji: 14349,

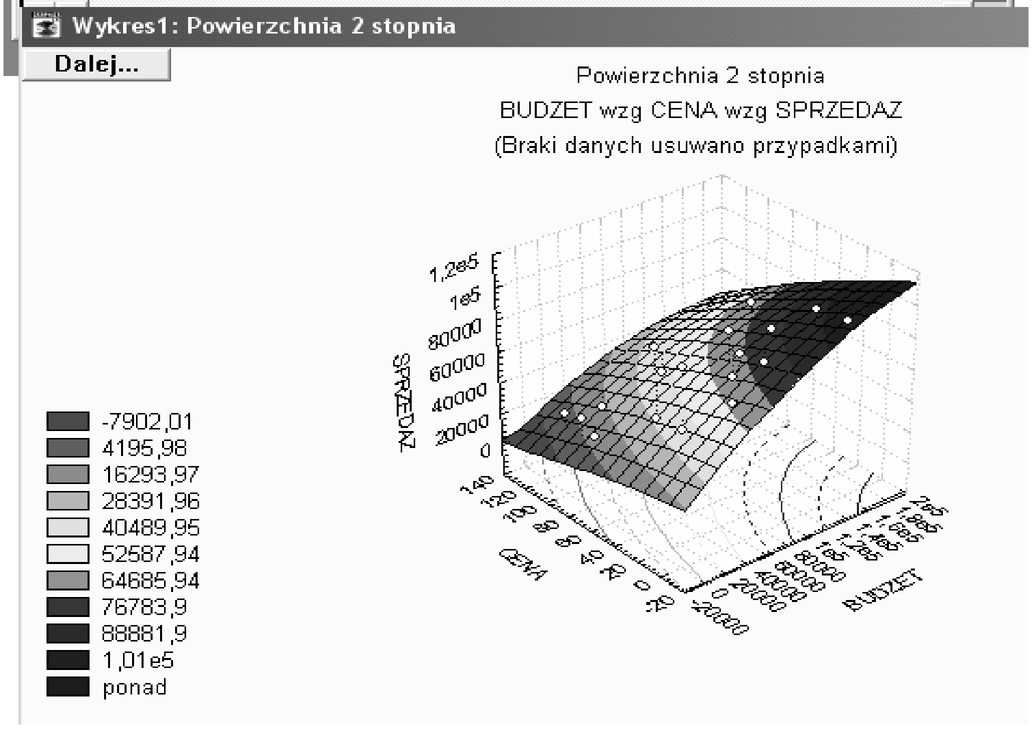
N=18	BETA	Błąd st. BETA	B	Błąd st. B	t(15)	poziom
W. wolny			36779,49	13165,54	2,79362	,0136
BUDZET	,593322	,144812	,38	,09	4,09720	,0009
CENA	-,400001	,144812	-358,14	129,66	-2,76222	,0145


Korelacje (podypl3.sta)

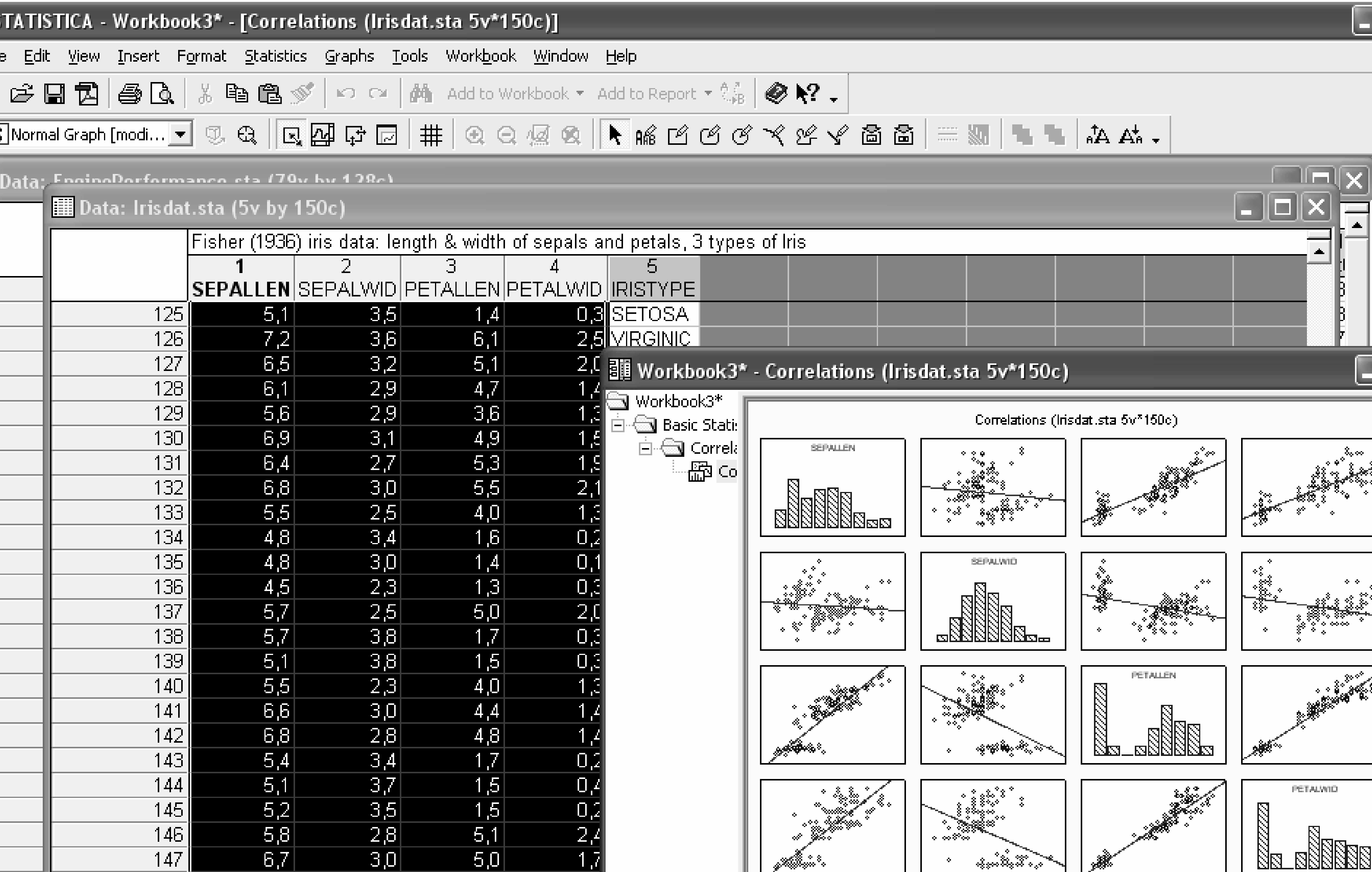
Dalej...

Oznaczone wsp. korelacji są istotne z p < ,05000
N=18 (Braki danych usuwano przypadkami)

Zmienna	BUDZET	CENA	SPRZEDAZ
BUDZET	1,00	-,62	,84
CENA	-,62	1,00	-,77
SPRZEDAZ	,84	-,77	1,00



Correlation matrix – analyse inluence of variables



Non-linear regression and linear transformations

- In many cases relationships between y and x are not linear but more complicated ones.
- Several techniques for solving the non-linear forms.
- Def.: Model *Y* = *f*(*X*,*b*) is linear with respect to parameters, if it possible to present it as a *linear* function of *univocal transformations of X*.

$$Y = \sum_{i=1}^{k} b_k z_k \qquad Z_k = h_k(X)$$

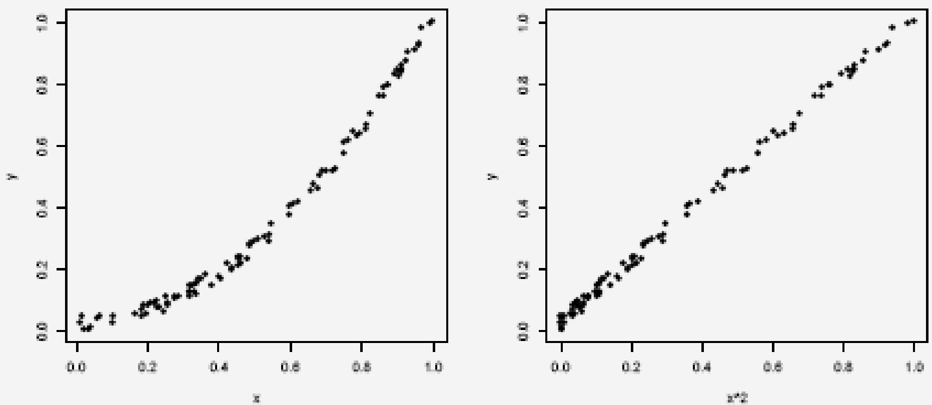
Transformations

- Why might we want to transform the variables?
- If the linear model appears to be quite poor, then transforming y can often improve the model's fit. Basically, transforming the data can linearize non-linear relationships.
- If some of the required conditions are violated, then transforming the variable can solve the problems.
- **No information is lost in a transformation**, but care must be taking in interpreting the coefficients, and the transformed model must be validated.
- Popular transformations:
  - Logarithms
  - Quadratic functions
  - Hyperbolic functions

Simple transformations

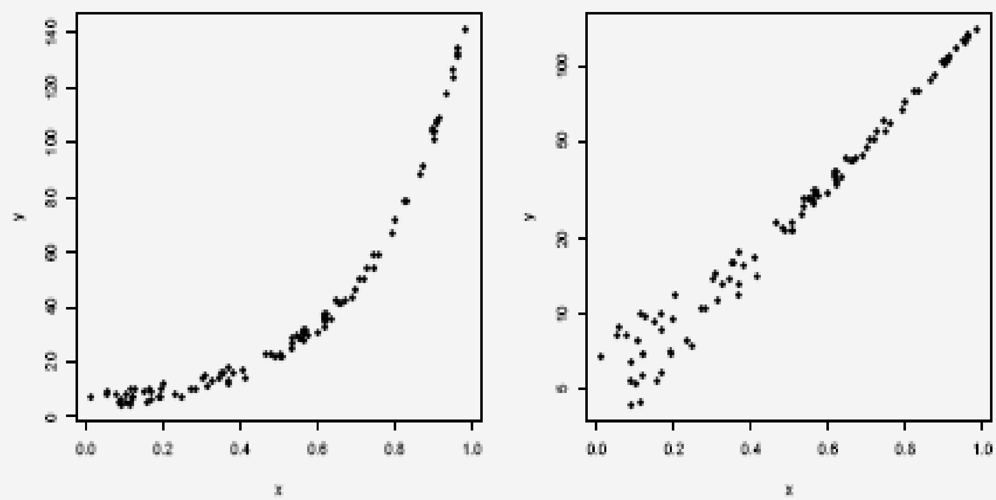
Co dają nam transformacje wielomianowe?

$$x' = x^2$$



logarytmowanie?

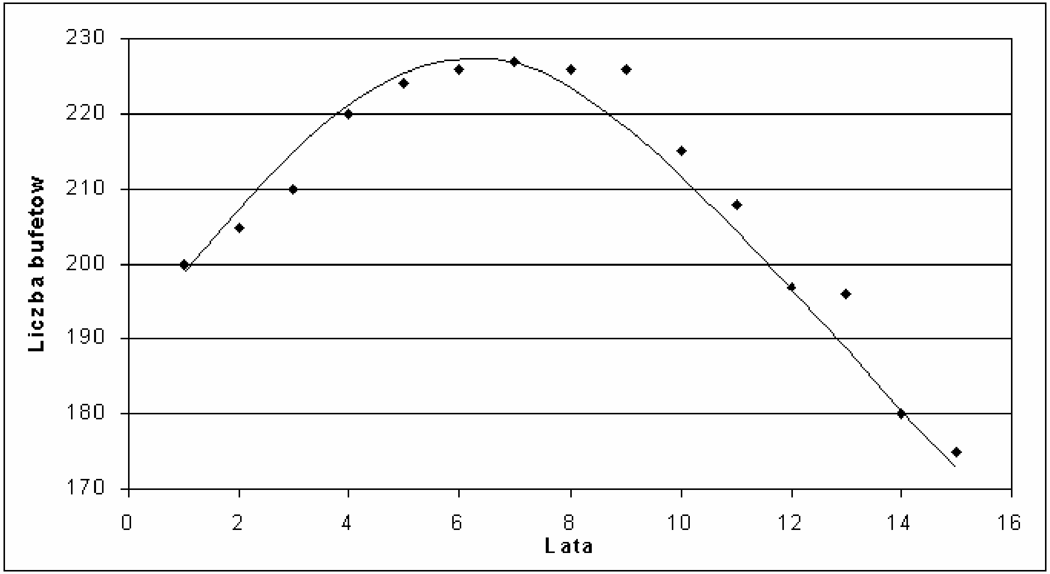
$$y' = log(y)$$



A toy example for quadratic functions

- Punkty żywieniowe w latach 1981-1995 (bars, cantines, …)

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



Toy examples again

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 $z_1 = t$ $z_2 = t^2$
- Rozwiązanie
  - a0=188
  - a1=11,031
  - a2=-0,814
- Weryfikacja
  - R2=0.996 s=3,37
  - Obie wartości statystyk *t* < *0.05*

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

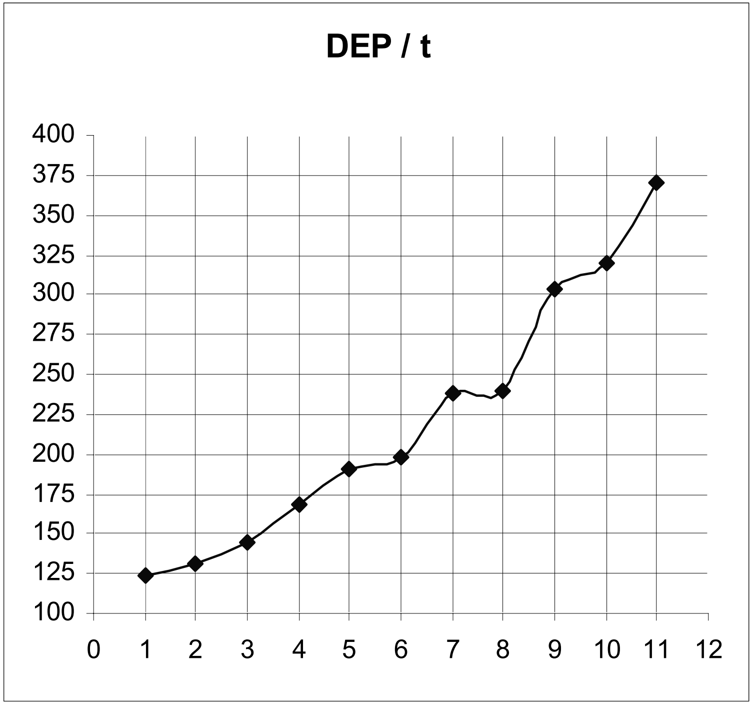
Logarithmic or exponential transformations

- Recall $y = e^a$.
- Recall $\log_e(y) = \ln y = a$. The natural log of a number is the power of *e* that produces that number.

Logarithmic transformations

- 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

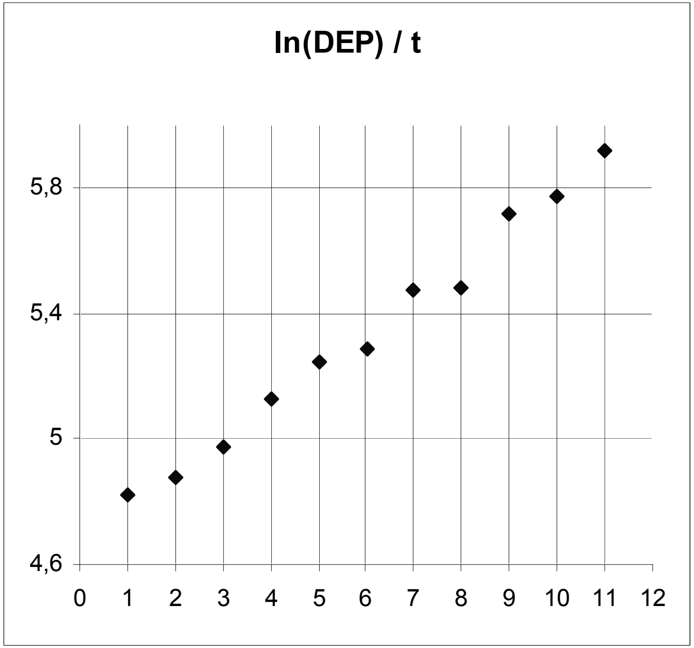


Hipoteza – wykładniczy przebieg $DEP = a \cdot e^{b \cdot t}$

Logarithmic transformations

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

Logarithmic transformations

- Rozwiązanie modelu przekształconego

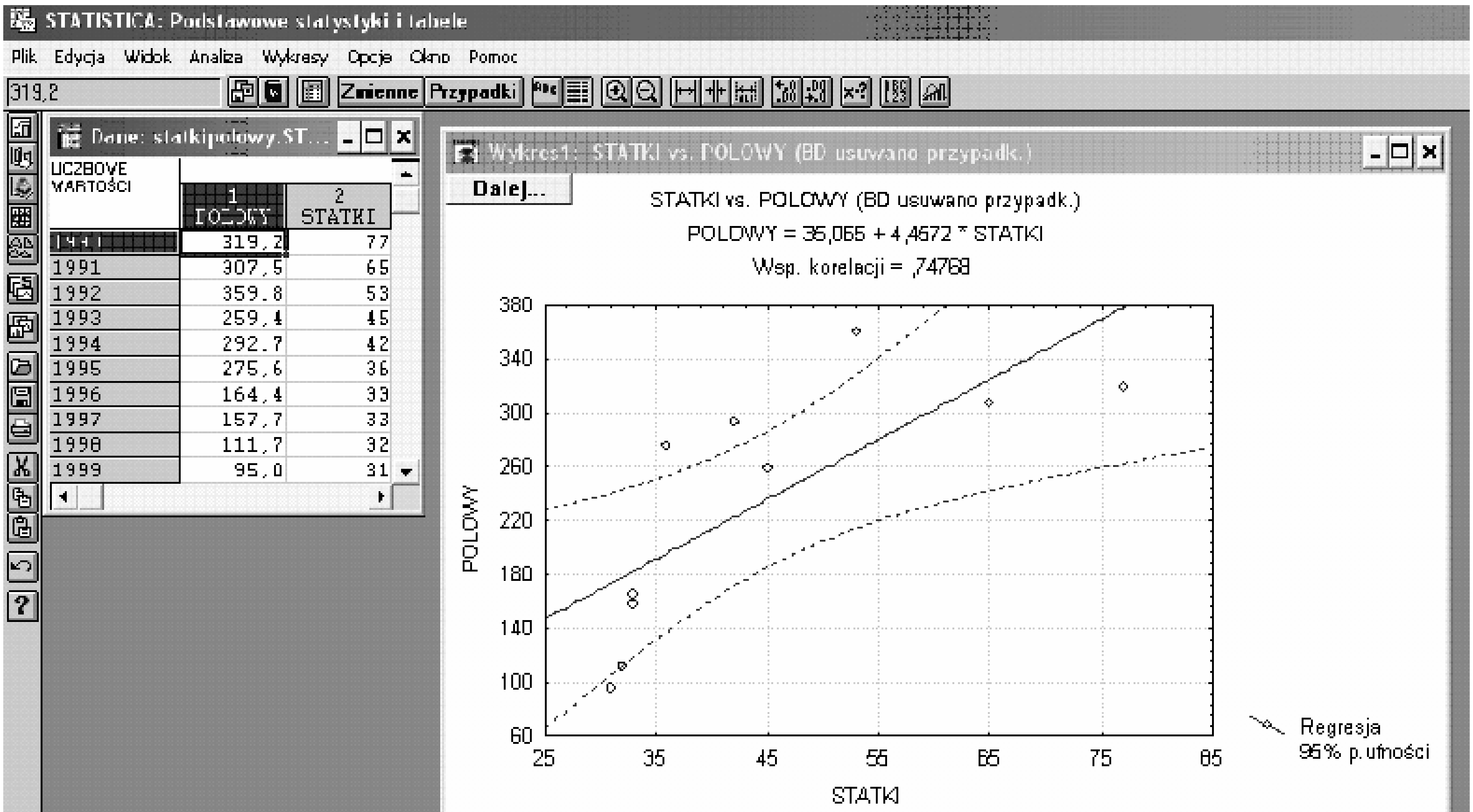
  ln(*DEP*)=4.671+0.111·*t*, R2=0.989, współczynniki istotne.

- Przekształcenie odwrotne

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

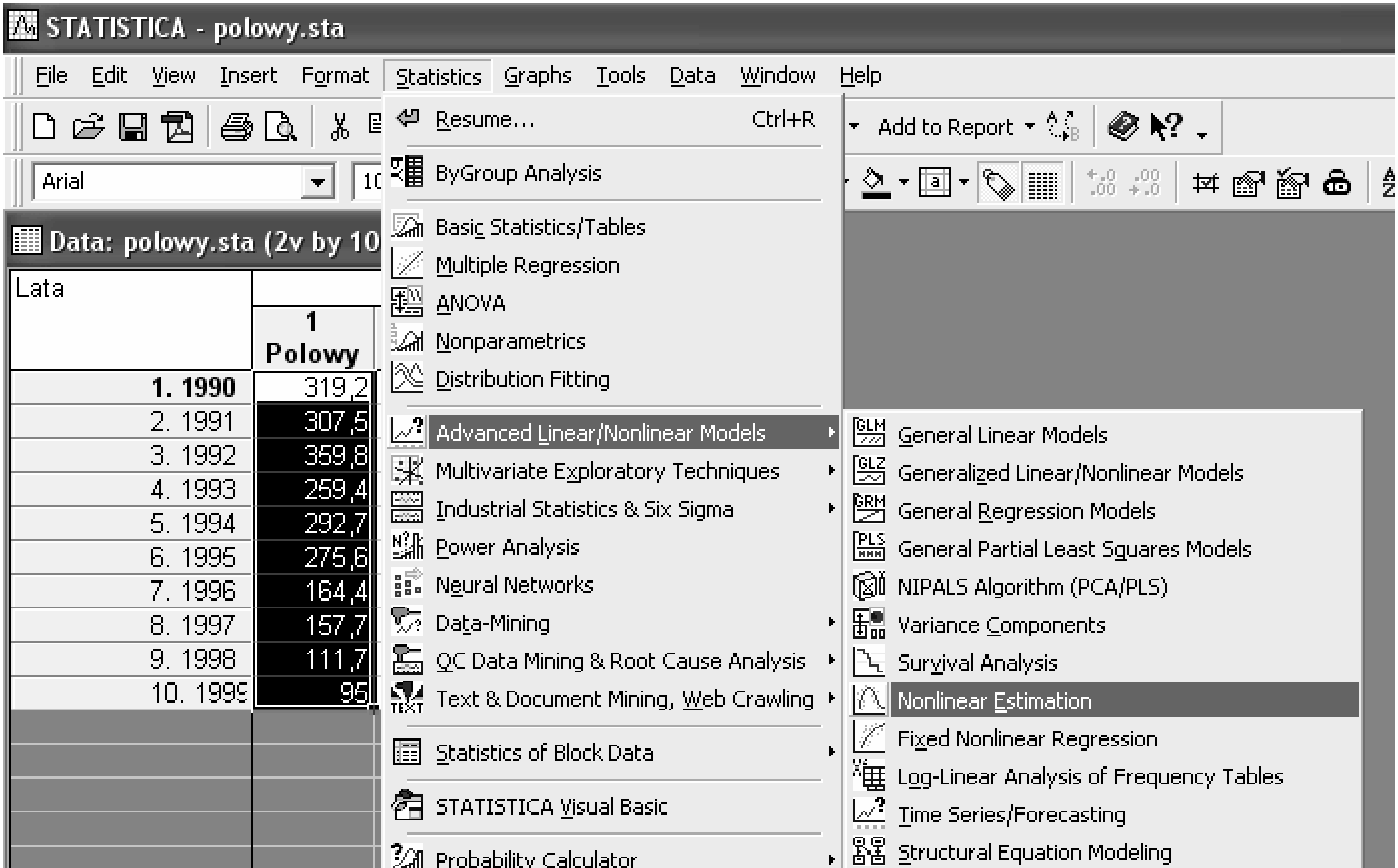
Nonlinear regression function

- Dane nt. polskiego rybołówstwa dalekomorskiego (lata 90te).
- Overseas fishermen / ships



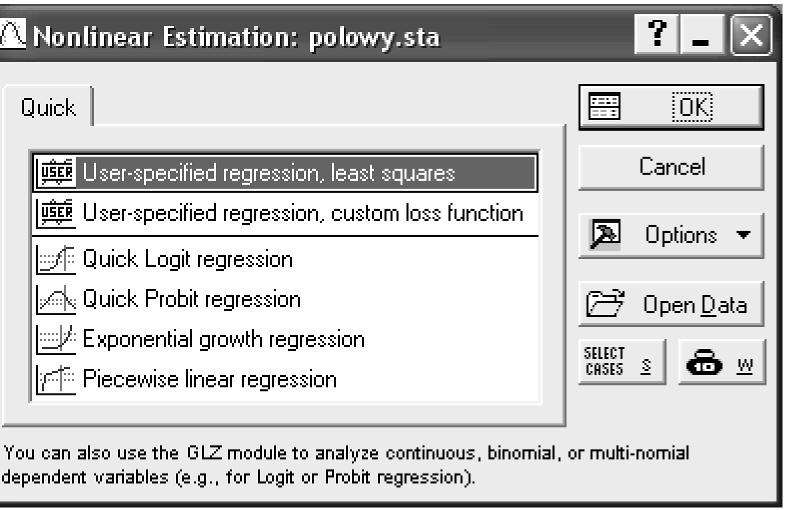
Statistica – example of nonlinear estimation

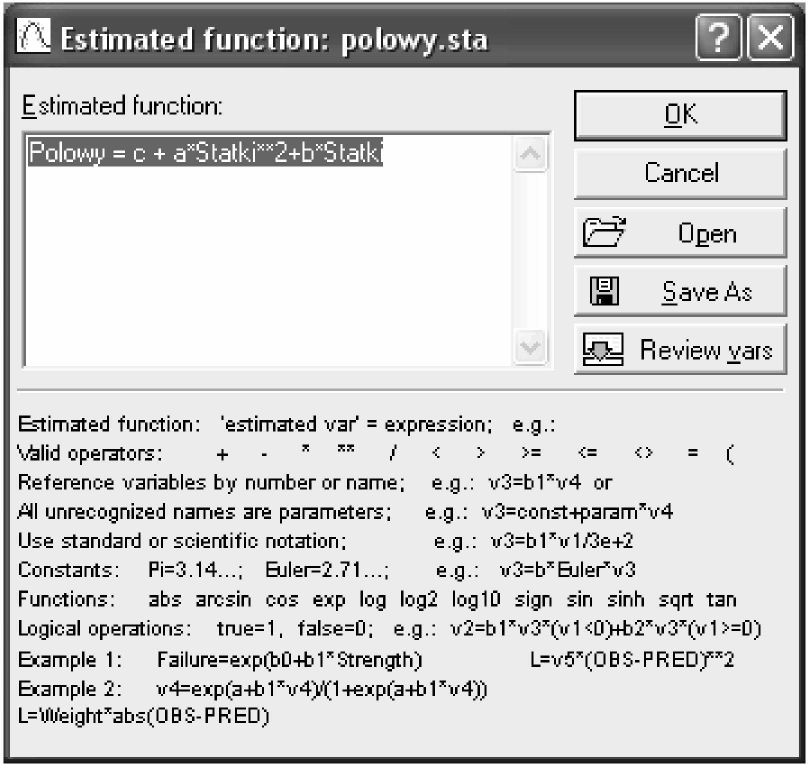
- Where to find it?



Fishing ships again

- User defined quadratic function

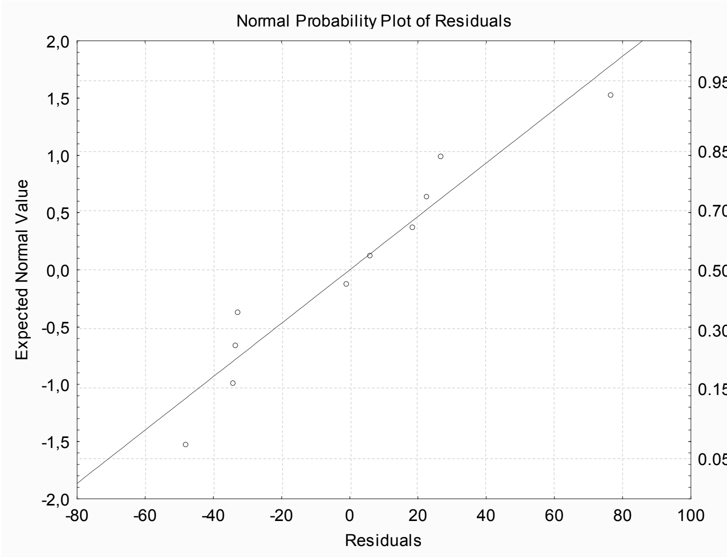




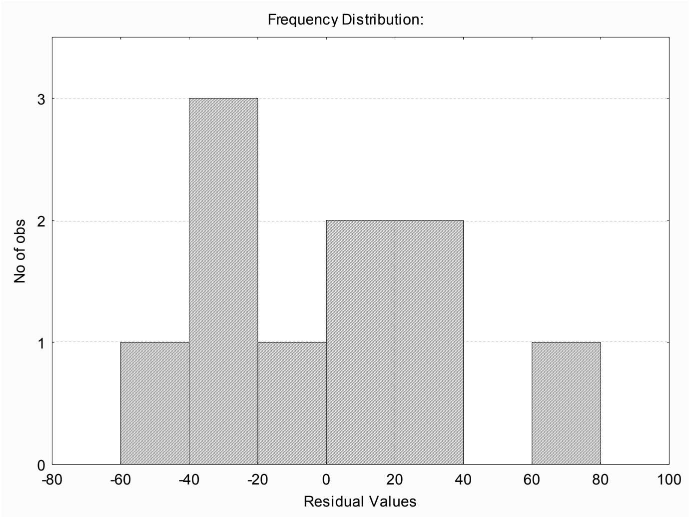
And some results

- Main reports

	Model is: Polowy = c + a*Statki**2+b*Statki (polowy.sta) Dep. Var. : Polowy Level of confidence: 95.0% (alpha=0.050)					
	Estimate	Standard error	t-value df = 7	p-level	Lo. Conf Limit	Up. Conf Limit
c	-581,494	185,4557	-3,13549	0,016483	-1020,03	-142,961
a	-0,251	0,0733	-3,41842	0,011159	-0,42	-0,077
b	30,708	7,7336	3,97071	0,005387	12,42	48,995



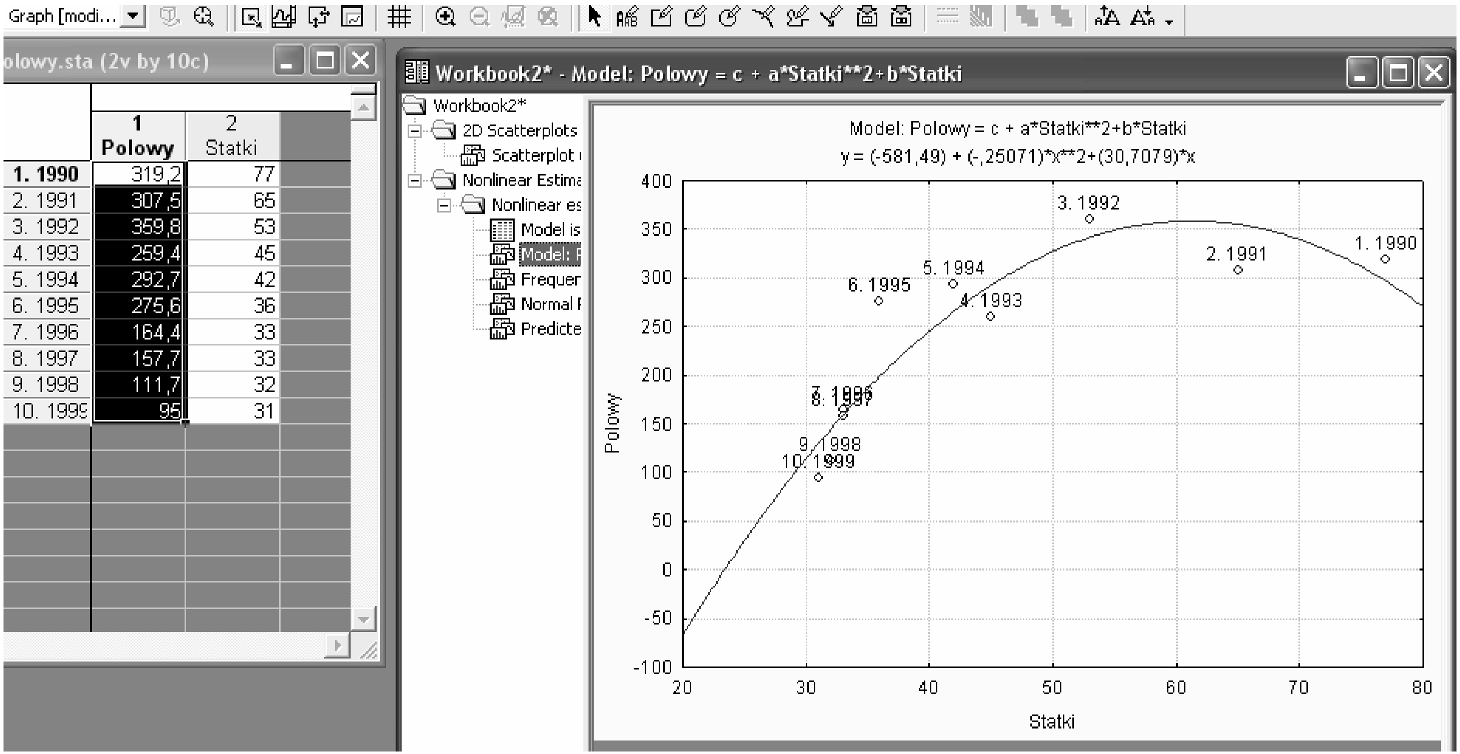
	Model is: Polowy = c + a*Statki**2+b*Statki (polowy.sta) Dep. Var. : Polowy				
Effect	1 Sum of Squares	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	615124,3	3,00000	205041,4	109,6297	0,000003
Residual	13092,2	7,00000	1870,3		
Total	628216,5	10,00000			
Corrected Total	79251,6	9,00000			
Regression vs.Corrected Total	615124,3	3,00000	205041,4	23,2850	0,000141



Non-linear regression

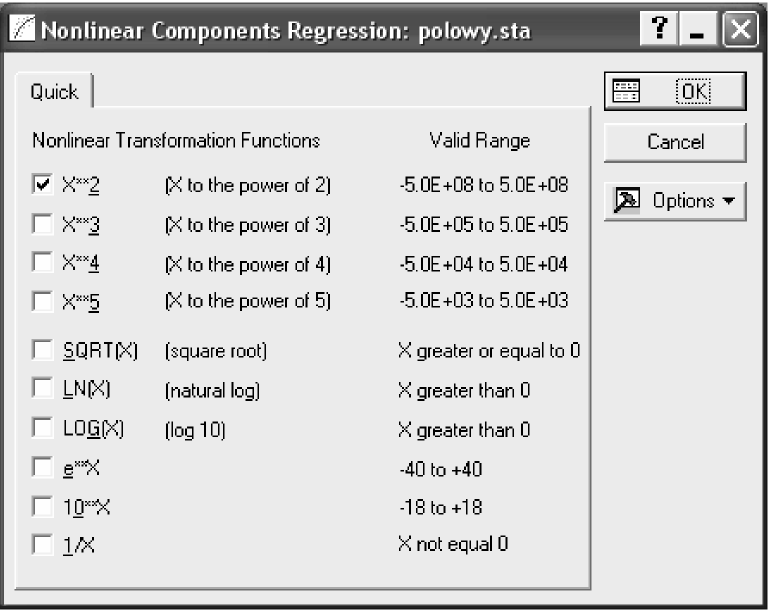
- Model funkcji kwadratowej (a co z innymi?)

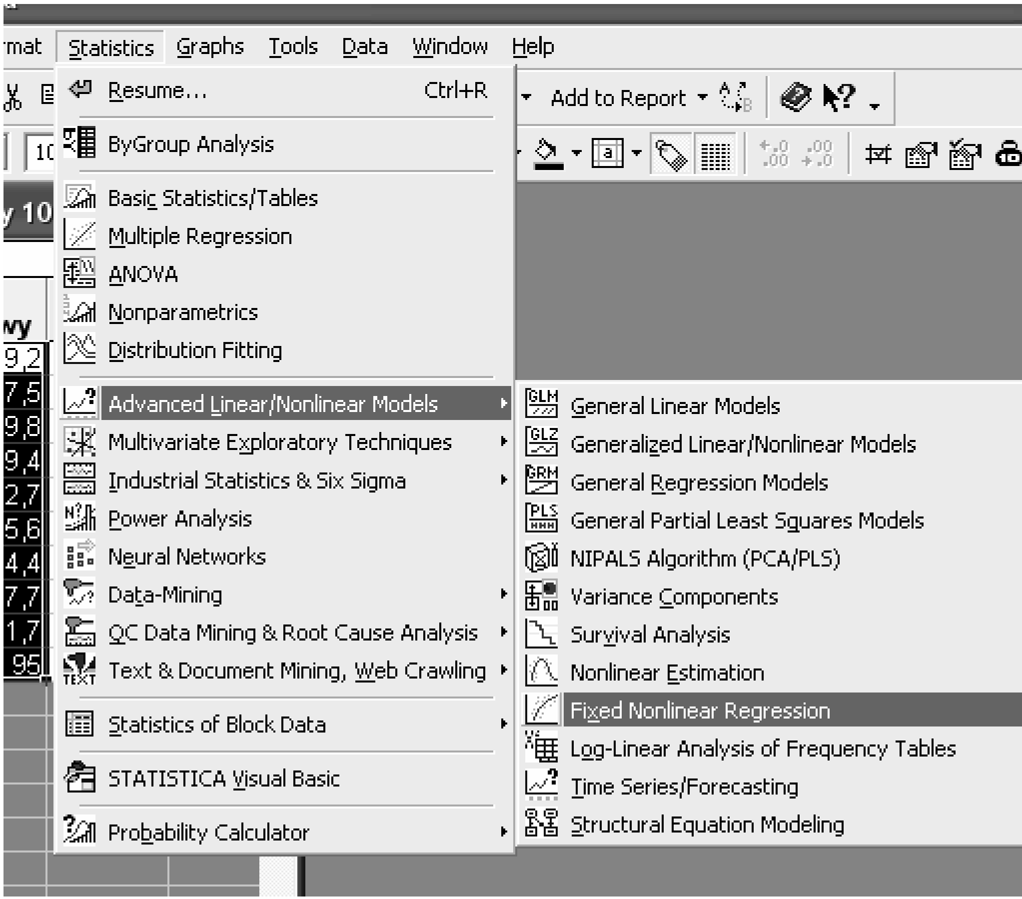
$$y = -0{,}25071 \cdot x^2 + 30{,}7079 \cdot x - 581{,}49$$



Statistica – another solution

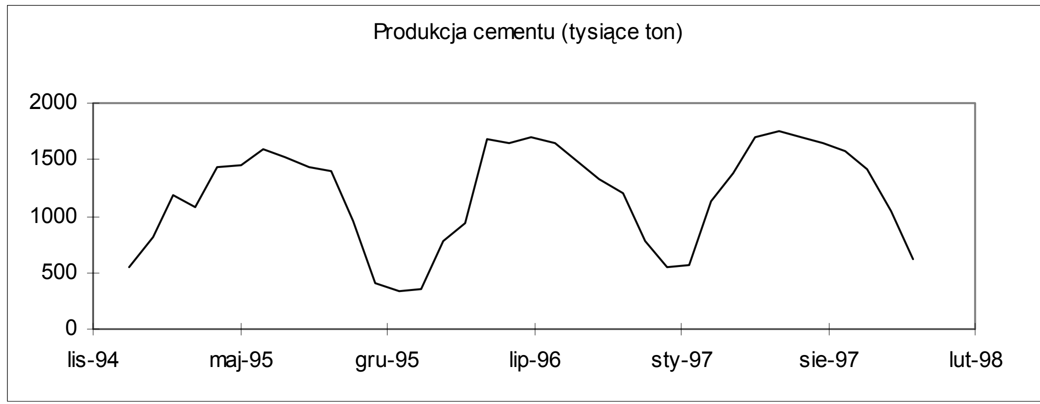
- Fitting fixed nonlinear basic functions





More difficult non-linear function

- Linear transformation – not always applicable → some cases are too difficult



- Non-parametric regression → approximate form is unknown for a user with respect to parameters
- Przeczytaj więcej w rozdziale 5tym książki Koronacki, Ćwiek „Statystyczne systemy uczące się” wyd. 2

Marginal piecewise functions

- Try to analyse local regions of data - segementation
- Combine (Add) several base function approximating local segments (the goal is to fit the data with a broken line)
- Regresyjne funkcje sklejane z węzłami (functions with knots)

Locally weighted regression

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

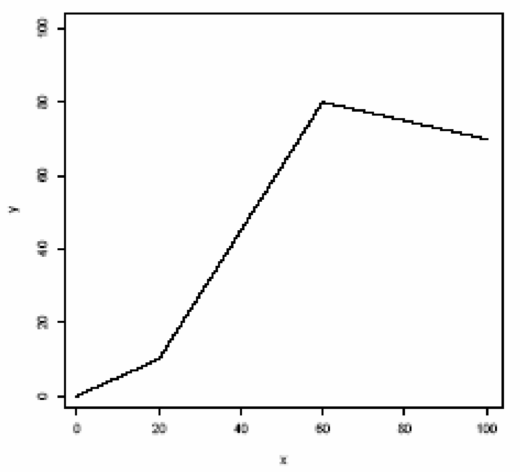


Figure 1.1. An Illustration of Linear Regression Splines with Two Knots

Read more on advanced regression …

- Comprehensive review in „The Data Mining and Knowledge Discovery Handbook” O.Maimon, L. Rokach (eds), Springer 2005.
- Analyse 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.

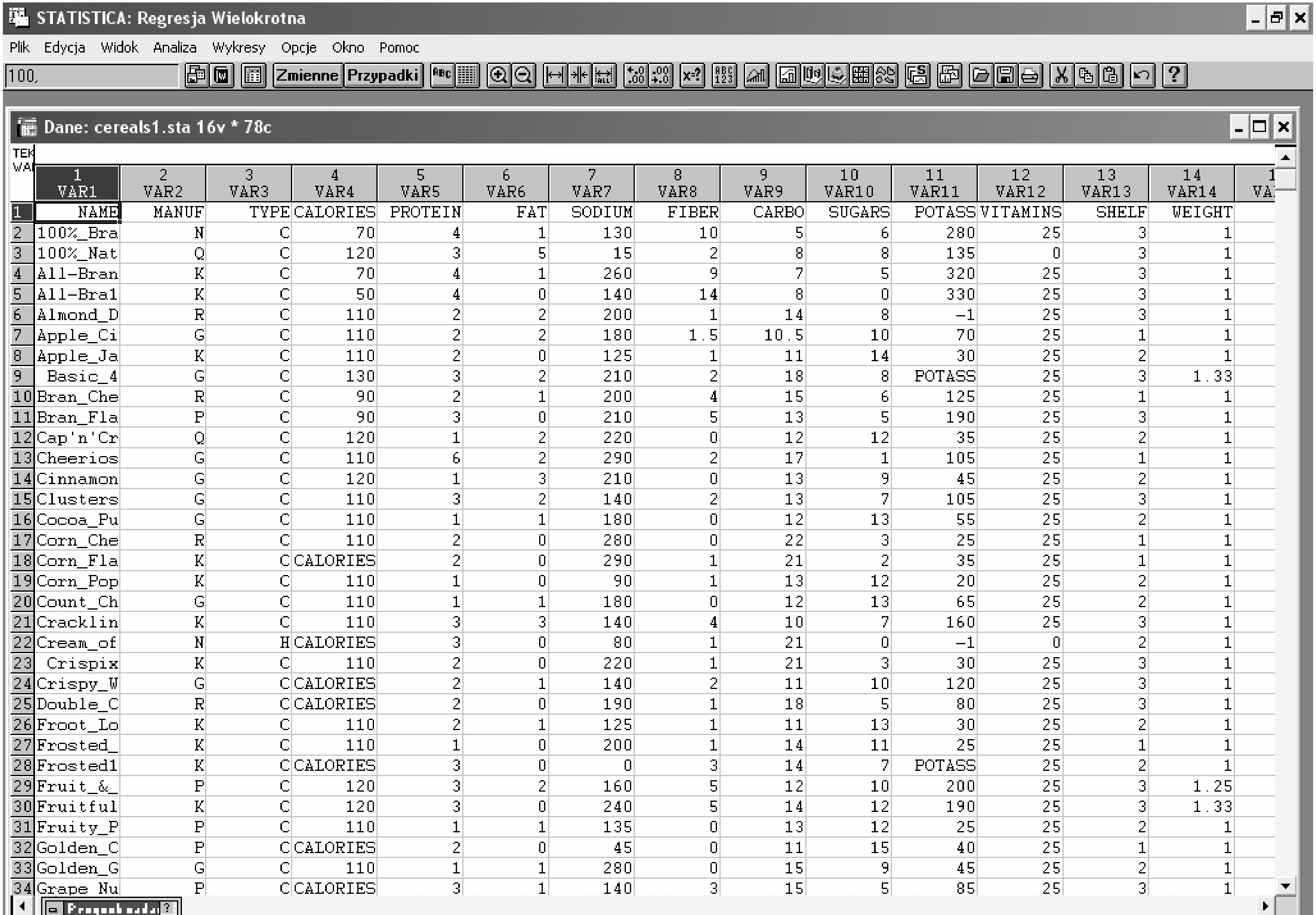
There is no requirement in regression analysis for there to be a "model" by which the data were supposed to be generated. There is no need to address cause and effect. And there is no need to undertake statistical tests or construct confidence intervals. The definition of a regression analysis can be 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.

Selecting Variables to Regression Model

- While building multiple regression:
- Should we use all available predictors (independent variables)
- Several approaches to select the subset of variables
- Let us remind local test in diagnostics of the model

Cereals examples from Larose's book



	1 VAR1	2 VAR2	3 VAR3	4 VAR4	5 VAR5	6 VAR6	7 VAR7	8 VAR8	9 VAR9	10 VAR10	11 VAR11	12 VAR12	13 VAR13	14 VAR14
1	NAME	MANUF	TYPE	CALORIES	PROTEIN	FAT	SODIUM	FIBER	CARBO	SUGARS	POTASS	VITAMINS	SHELF	WEIGHT
2	100%_Bra	N	C	70	4	1	130	10	5	6	280	25	3	1
3	100%_Nat	Q	C	120	3	5	15	2	8	8	135	0	3	1
4	All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1
5	All-Bra1	K	C	50	4	0	140	14	8	0	330	25	3	1
6	Almond_D	R	C	110	2	2	200	1	14	8	-1	25	3	1
7	Apple_Ci	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1
8	Apple_Ja	K	C	110	2	0	125	1	11	14	30	25	2	1
9	Basic_4	G	C	130	3	2	210	2	18	8	POTASS	25	3	1.33
10	Bran_Che	R	C	90	2	1	200	4	15	6	125	25	1	1
11	Bran_Fla	P	C	90	3	0	210	5	13	5	190	25	3	1
12	Cap'n'Cr	Q	C	120	1	2	220	0	12	12	35	25	2	1
13	Cheerios	G	C	110	6	2	290	2	17	1	105	25	1	1
14	Cinnamon	G	C	120	1	3	210	0	13	9	45	25	2	1
15	Clusters	G	C	110	3	2	140	2	13	7	105	25	3	1
16	Cocoa_Pu	G	C	110	1	1	180	0	12	13	55	25	2	1
17	Corn_Che	R	C	110	2	0	280	0	22	3	25	25	1	1
18	Corn_Fla	K	C	CALORIES	2	0	290	1	21	2	35	25	1	1
19	Corn_Pop	K	C	110	1	0	90	1	13	12	20	25	2	1
20	Count_Ch	G	C	110	1	1	180	0	12	13	65	25	2	1
21	Cracklin	K	C	110	3	3	140	4	10	7	160	25	3	1
22	Cream_of	N	H	CALORIES	3	0	80	1	21	0	-1	0	2	1
23	Crispix	K	C	110	2	0	220	1	21	3	30	25	3	1
24	Crispy_W	G	C	CALORIES	2	1	140	2	11	10	120	25	3	1
25	Double_C	R	C	CALORIES	2	0	190	1	18	5	80	25	3	1
26	Froot_Lo	K	C	110	2	1	125	1	11	13	30	25	2	1
27	Frosted_	K	C	110	1	0	200	1	14	11	25	25	1	1
28	Frosted1	K	C	CALORIES	3	0	0	3	14	7	POTASS	25	2	1
29	Fruit_&_	P	C	120	3	2	160	5	12	10	200	25	3	1.25
30	Fruitful	K	C	120	3	0	240	5	14	12	190	25	3	1.33
31	Fruity_P	P	C	110	1	1	135	0	13	12	25	25	2	1
32	Golden_C	P	C	CALORIES	2	0	45	0	11	15	40	25	1	1
33	Golden_G	G	C	110	1	1	280	0	15	9	45	25	2	1
34	Grape_Nu	P	C	CALORIES	3	1	140	3	15	5	85	25	3	1

Summary of regression

- Analyse significance of each parameter
- Local test

,363690376281738 Kolumny Wiersze

Dalej...

R= ,55132922 R2= ,30396391 Popraw. R^2= ,20007793
F(10,67)=2,9259 p<,00422 Błąd std. estymacji: 35,095

N=78	BETA	Błąd st. BETA	B	Błąd st. B	t(67)	poziom p
W. wolny			-81,1243	29,74352	-2,72746	,008140
VAR4	,38579	,140748	,7814	,28508	2,74100	,007845
VAR5	-1,04006	1,137198	-3,6807	4,02440	-,91458	,363690
VAR6	-2,75604	1,412080	-9,6106	4,92410	-1,95176	,055149
VAR7	-,13467	,118025	-,0632	,05542	-1,14102	,257923
VAR8	-,01215	,222610	-,0179	,32713	-,05458	,956633
VAR9	,12430	,218372	,2241	,39372	,56921	,571118
VAR10	-,13873	,339894	-,4764	1,16719	-,40814	,684469
VAR11	,32140	,112291	,1781	,06221	2,86218	,005609
VAR12	-,11344	,129004	-,1883	,21416	-,87934	,382358
VAR13	4,14593	1,446740	14,6513	5,11264	2,86570	,005554

Variable selection

- Choose variables according to domain knowledge
- Necessary properties:
  - Should be correlated with target y.
  - Dependent variables x cannot be correlated (or at least not highly ..).
  - As to domain of x – high variability.
- Use heuristics on typical correlation coefficient

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

Toy example

- 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	[illegible]	[illegible]	0,4	[illegible]	0,05	[illegible]

Toy example – basic compuations

- Współczynniki zmienności – Standarized Variability

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

- Correlation matrix

	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

Calculations again

- 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?

Stepwise regression

- Postępująca (*forward*)
  - Add one by one these variables which inluence target y in the highest way (with respect to evaluation measures).
- Wsteczna (*backward*)
  - Start with all variables, remove the weakest one.

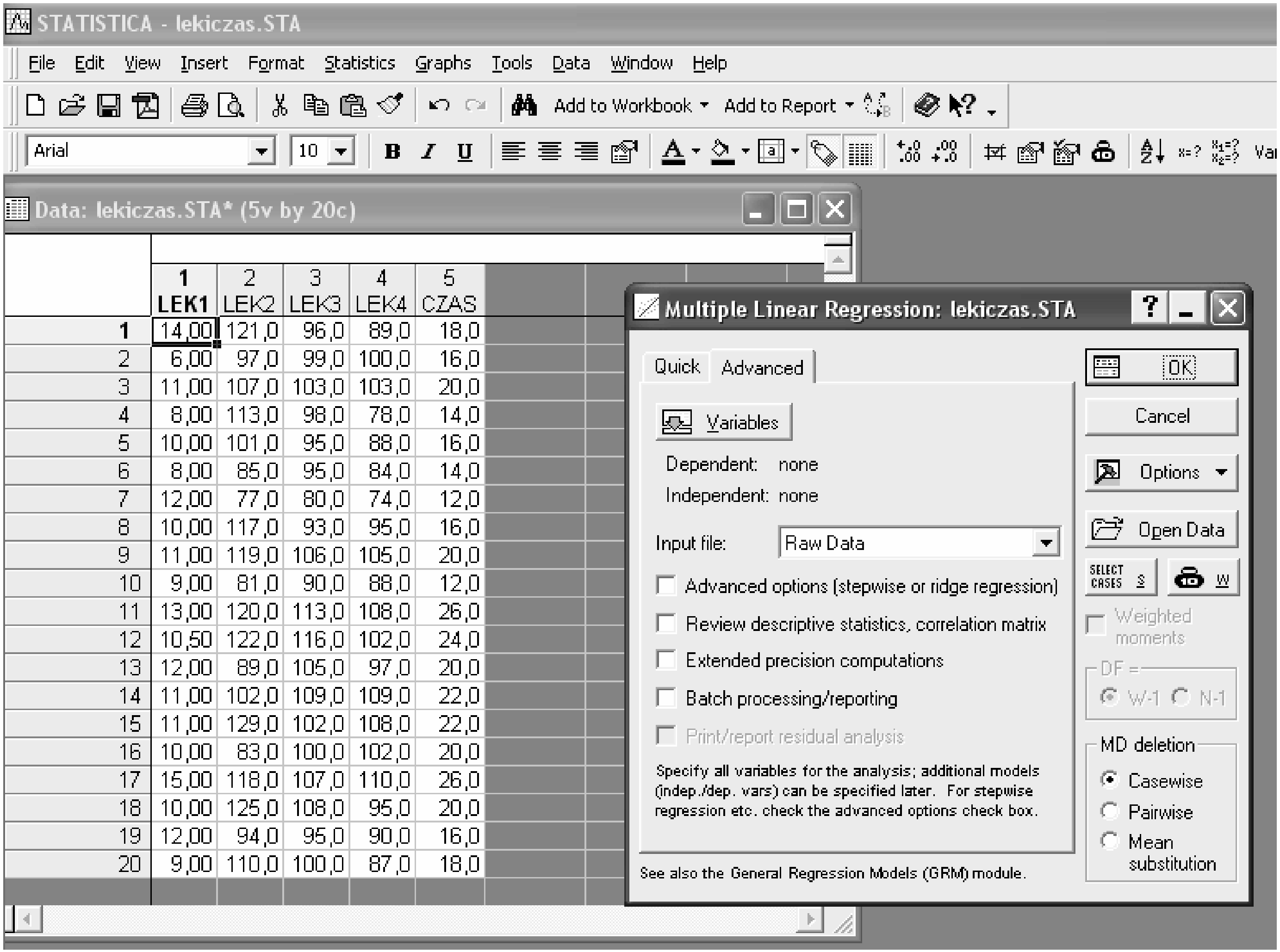
  - You can apply such evaluators as
  - Stosując R2 lub testy istotności współczynników modelu (*F*).

Stepwise regression with R2

- If R2 changes very little when a variable is added to the regression, then this is an indication that the additional variable doesn’t explain very much of the variation in *y*.
- If R2 changes a lot with the additional covariate, then this is just one indication a variable explains a lot of the variation in y.
- The difference between the R2 when the variation is excluded from the regression and the R2 when the variable is included in the regression can be used as a measure of how much variation in *y* that particular variable explains.

An example of forward selection – medicine 1

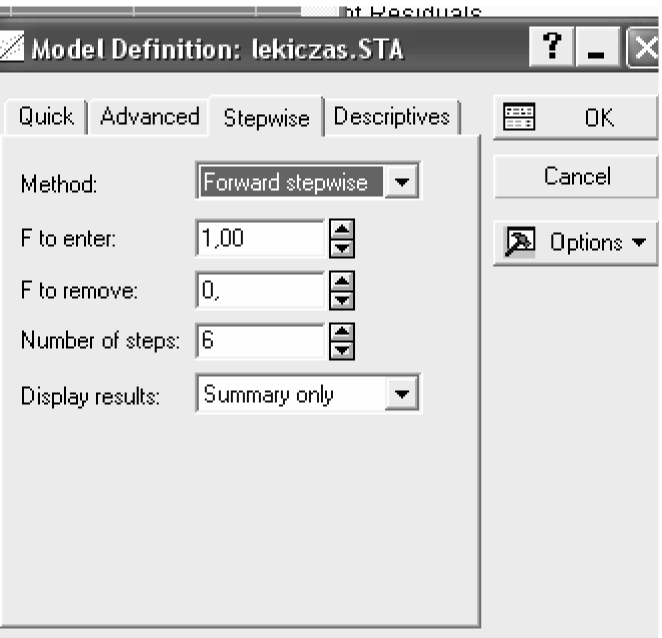
- Predict the length of patient's stay in a hospital depending on prescribed medicines.

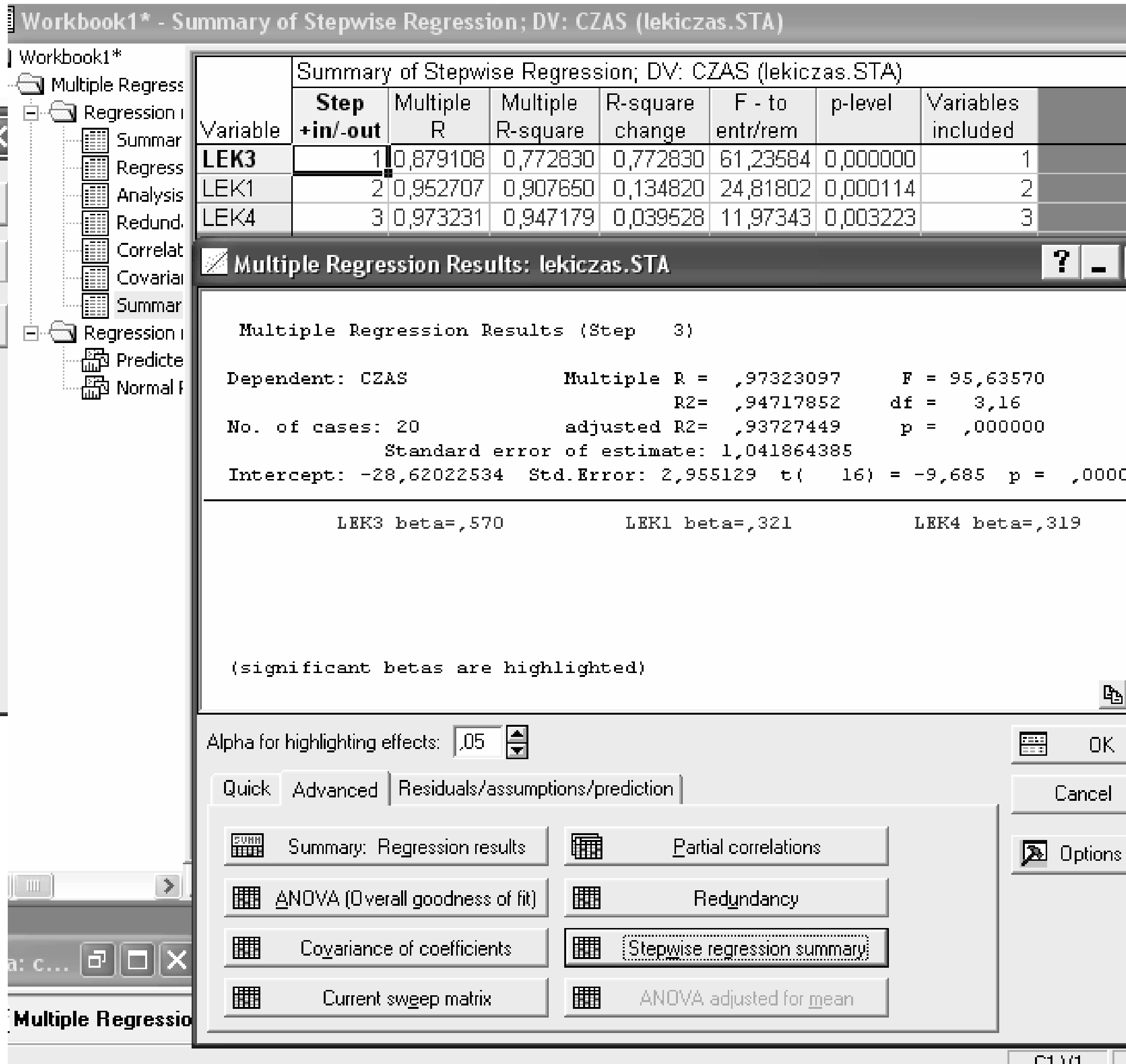


	1 LEK1	2 LEK2	3 LEK3	4 LEK4	5 CZAS
1	14,00	121,0	96,0	89,0	18,0
2	6,00	97,0	99,0	100,0	16,0
3	11,00	107,0	103,0	103,0	20,0
4	8,00	113,0	98,0	78,0	14,0
5	10,00	101,0	95,0	88,0	16,0
6	8,00	85,0	95,0	84,0	14,0
7	12,00	77,0	80,0	74,0	12,0
8	10,00	117,0	93,0	95,0	16,0
9	11,00	119,0	106,0	105,0	20,0
10	9,00	81,0	90,0	88,0	12,0
11	13,00	120,0	113,0	108,0	26,0
12	10,50	122,0	116,0	102,0	24,0
13	12,00	89,0	105,0	97,0	20,0
14	11,00	102,0	109,0	109,0	22,0
15	11,00	129,0	102,0	108,0	22,0
16	10,00	83,0	100,0	102,0	20,0
17	15,00	118,0	107,0	110,0	26,0
18	10,00	125,0	108,0	95,0	20,0
19	12,00	94,0	95,0	90,0	16,0
20	9,00	110,0	100,0	87,0	18,0

An example of forward selection – medicine 2

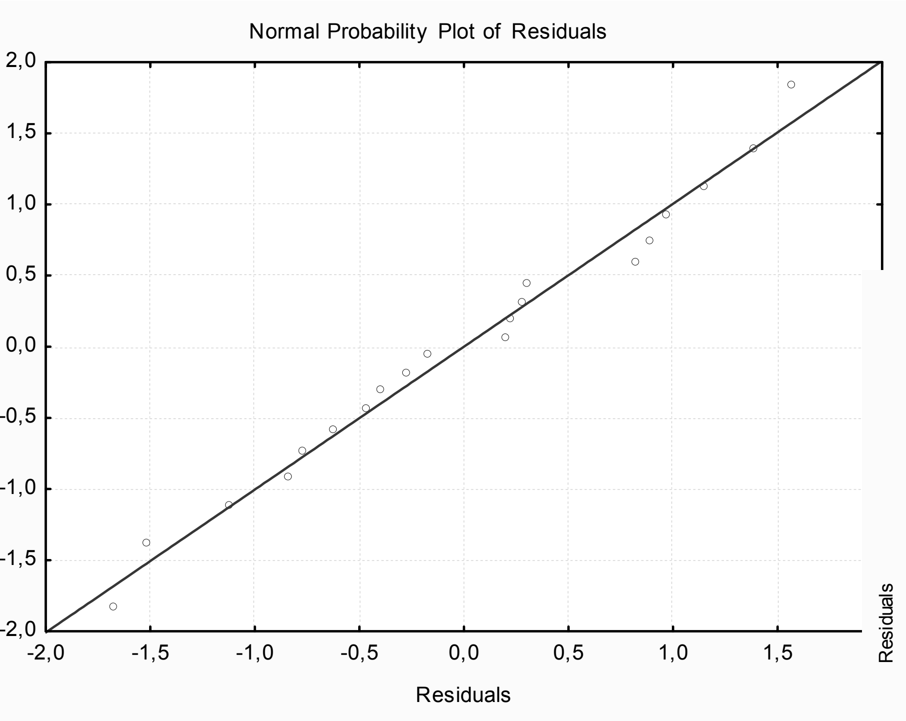
- Statsoft dialog and option windows

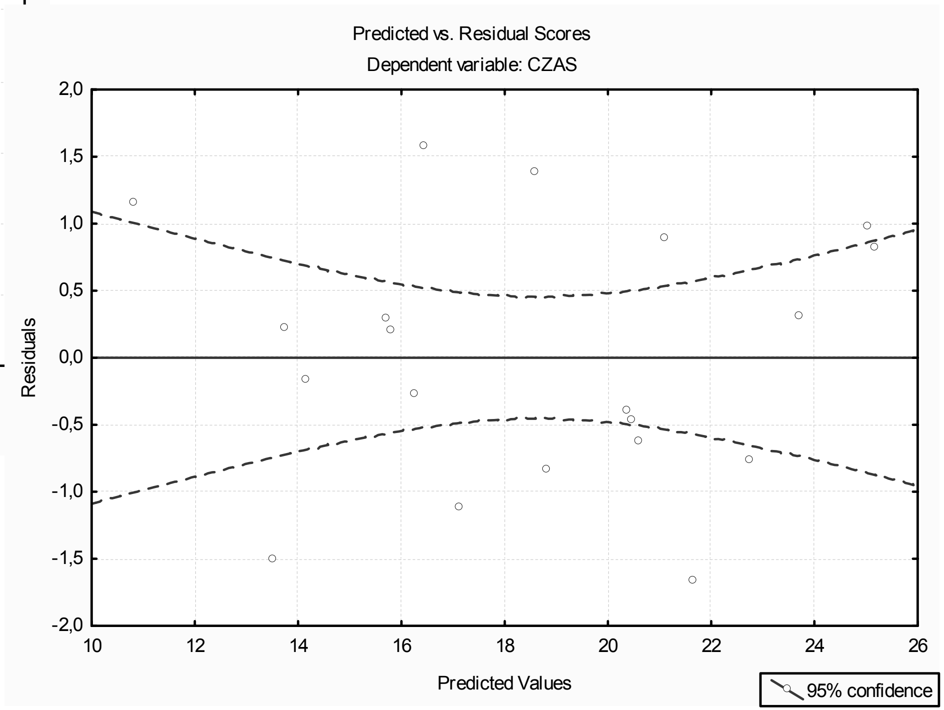




An example of forward selection – medicine 3

- Diagnostics of the linear model





Some comments to the example

- See *multicollinearity …*

Macierz korelacji

	LEK1	LEK2	LEK3	LEK4	CZAS
LEK1	1,00	,2729	,2019	,3169	,5371
LEK2	,2738	1,00	,602	,4556	,6068
LEK3	,20194	,60224	1,00	,76438	,8791
LEK4	,3169	,4558	,7644	1,00	,8569
CZAS	,5371	,60679	,8791	,8568	1,0

An example of regression in SE → COCOMO

- Cost estimate → a math. function of a number of the major cost factors:

$$Effort = f(x_1, x_2, ..., x_n)$$

- Cost factors → Boehm et al. COCOMO model.
- Functions:
  - Linear
  - Multiplicative
  - Power functions

$$E = a \cdot S^b \cdot F$$

S – code size; *a*,*b* functions of other cost factors

But: difficulties with parameterizations and no explanation of an influence of cost factors.

COCOMO cost factors:

- **Product factors**: required reliability, product complexity, database size used, required reusability, documentation match to life-cycle needs.
- **Computer factors**: execution time constraints, main storage constr., computer turnaround, platform volatility, type of computers.
- **Personel factors**: analyst capability, application experience, programming capability; platform experience; language and tool experience, personnel continuity.
- **Project factors**: multi-site development, use of soft. tools, required development schedule, …

Others, e.g. adaptation adjustment factors.

Example from Software Engineering

- Data of predicting costs (efforts) of producing softwer
- COCOMO database (one of the historical ones – after G.Ruhe study)
  - 63 historical projects described by 23 variables (concering - predictive COCOMO factors) and 2 outputs:
  - *tkdsi* - (size) total delivered source instructions, czyli rozmiaru dostarczonego ostatecznie kodu programu,
  - *effort* - effort in men/month, czyli ostateczny osobo-koszt projektu (pracochłonność).
- Literature function for COCOMO model

$$effort = a \cdot tkdsi^{b} \cdot eaf$$

Independent variables

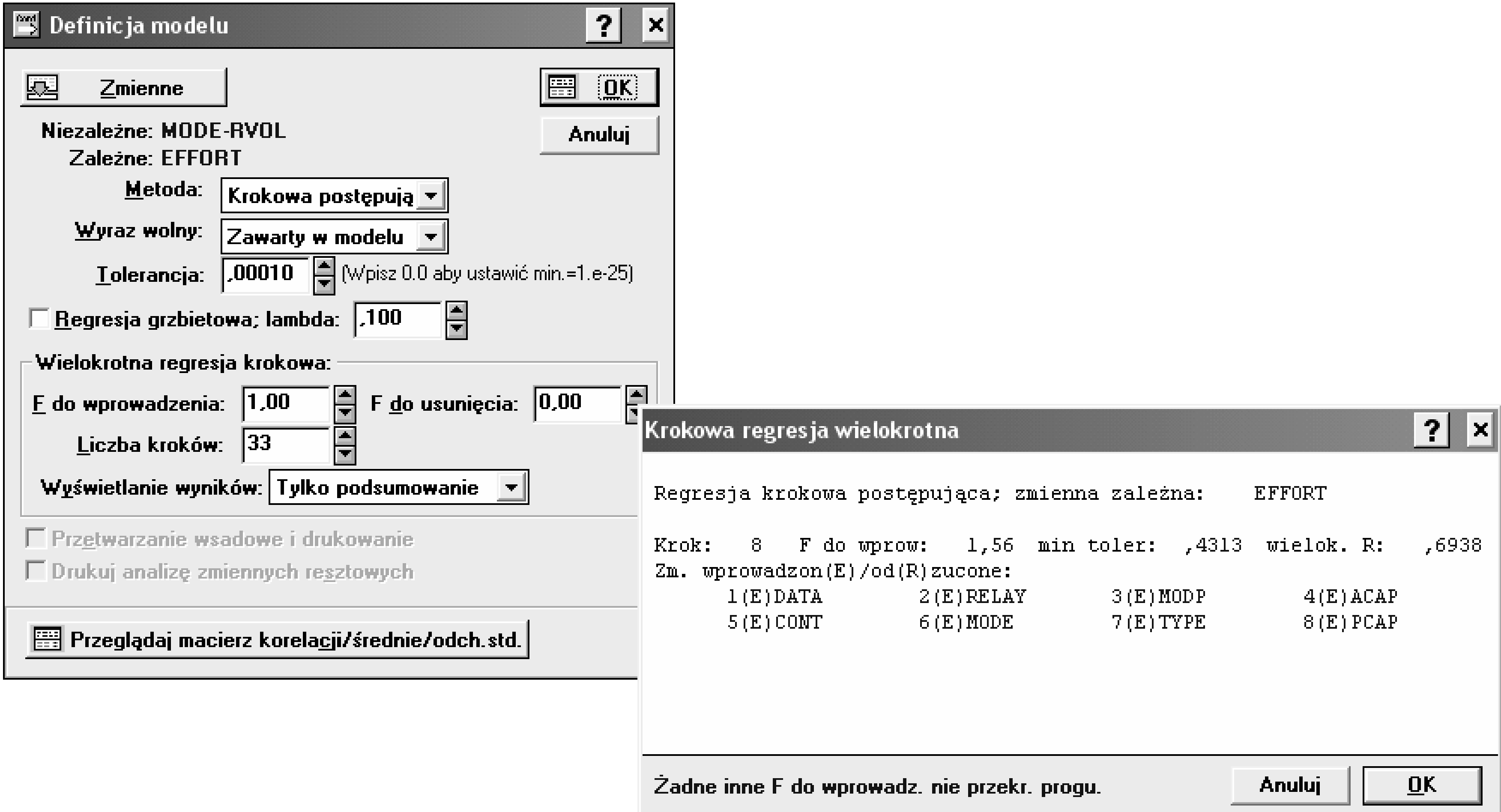
- mode - (oth) software development mode [embedded, semidetached, organic]
- appl - (oth) type of application [business, control, human-machine, scientific, support, system]
- lang - (Prod) language level [Fortran, Cobol, PL1, Pascal, APL, PL/S, Jovial, C, CMS-2]
- rely - (Prod) required software reliability
- data - (Prod) size of data base
- cplx - (Prod) product complexity
- aaf - (oth) adaptation adjustment factor (adjustment factor for size instructions that were adopted instead of newly developed)
- time - (Comp) execution time constraint
- stor - (Comp) main storage constraint
- virt - (Comp) virtual machine volatility (virtual machine = hardware and software under which works analysed SW system)
- turn - (Comp) computer turnaround time
- type - (Comp) type of computer [maxi, midi, mini, micro]
- acap - (Pers) analyst capability
- aexp - (Pers) applications experience
- pcap - (Pers) programmer capability
- vexp - (Pers) virtual machine experience
- lexp - (Pers) language experience
- cont - (Pers) personnel continuity [low, nominal, high]
- modp - (Proj) use of modern programming practices
- tool - (Proj) use of software tools
- sced - (Proj) required development schedule
- rvol - (Proj) requirements volatility

Target variables

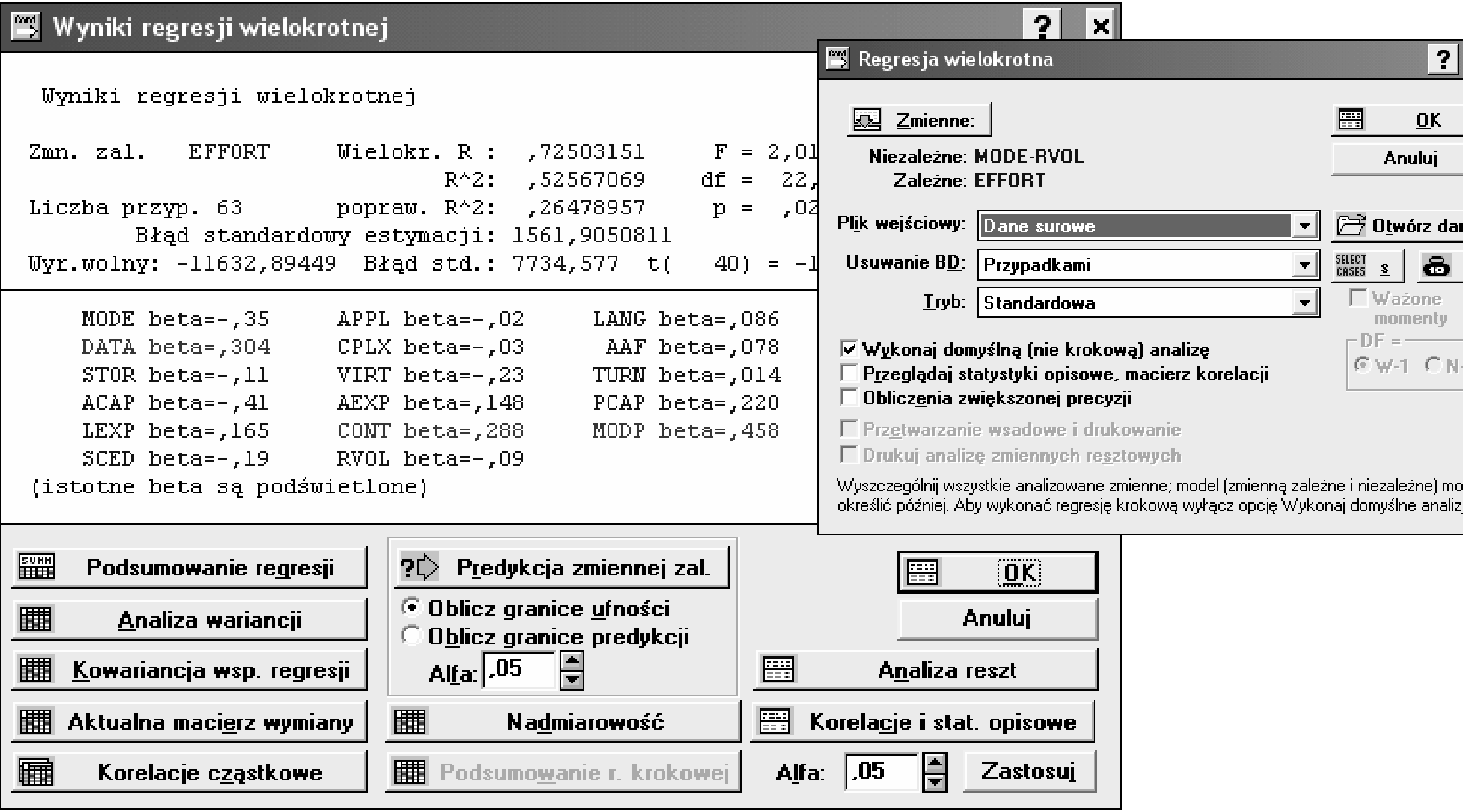
- What about the variable $$prod = \frac{tkdsi}{effort}$$
- Check residuals analysis
- Linear regression models (with all variables)

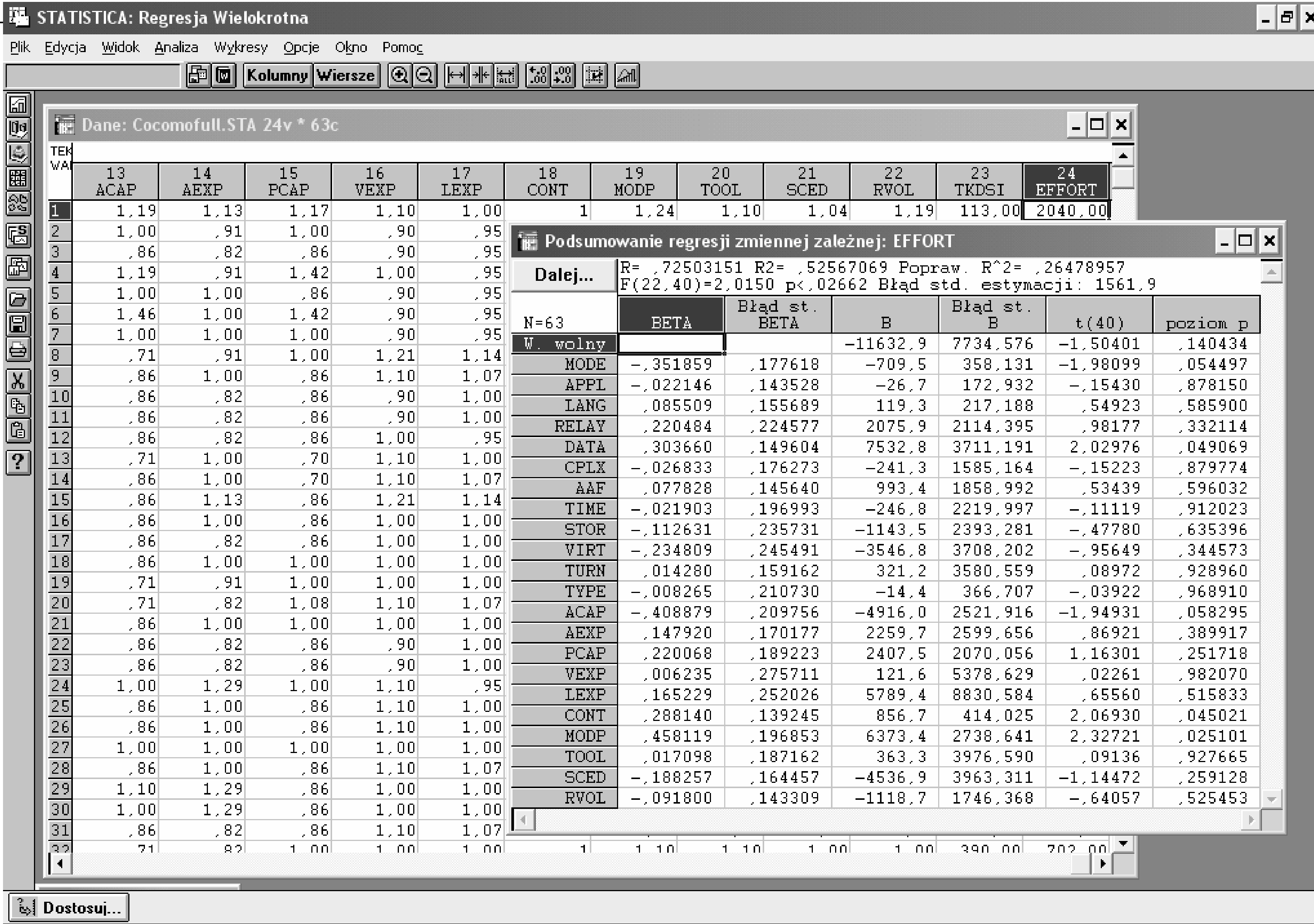
Zmienna	Współ determinacji R2
tkdsi	0.308
effort	0,422
log(tkdsi)	0.551
log(effort)	0.598
prod	0.721
log(prod)	0.906

Regresja krokowa – Stepwise regression



Multiple regression - 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

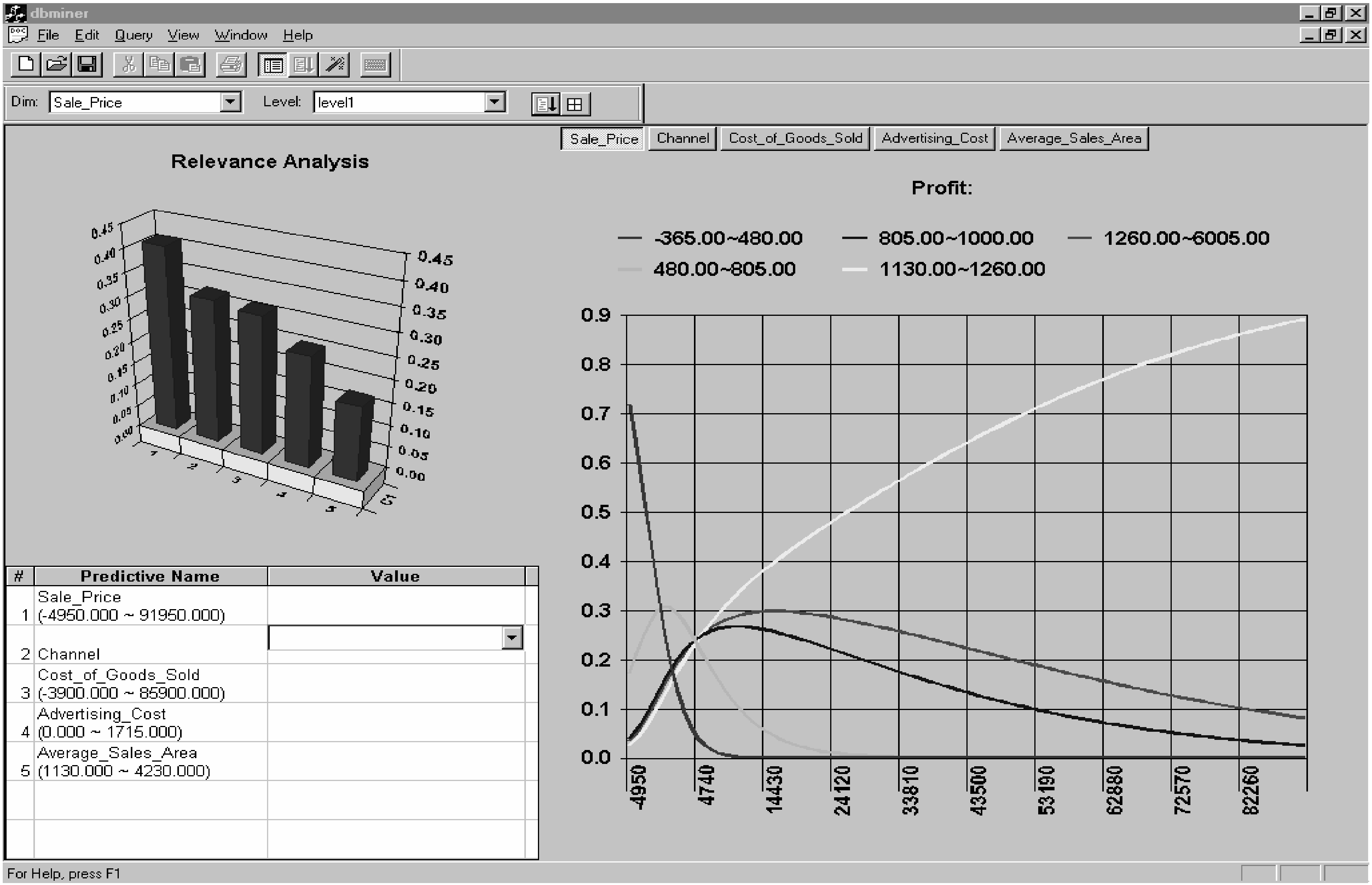
Dalej...

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...

Examples of Predicting Numerical Values



Applications – many, many …

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

Applications – few words more …

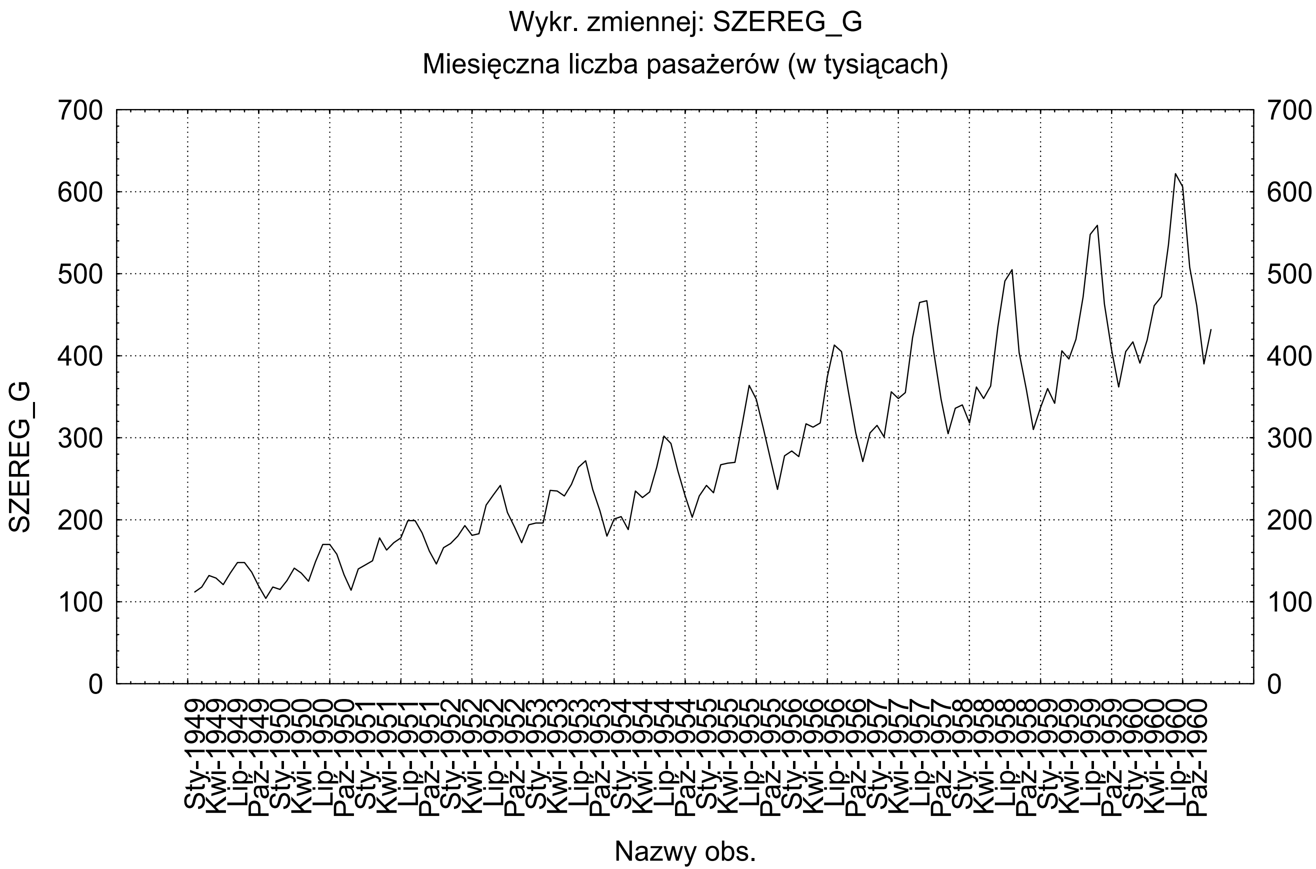
- Finance:
  - The capital asset pricing model uses linear regression as well as the other predictive models for analyzing and quantifying the systematic risk of an investment.
- Marketing
  - Analysis of sales, demands for products, …
- Economical sciences
  - Macro-economical models for countries, 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

Regression in time dependent data

- Observations are ordered with time stamps
- Typical example – time series
- Trend line
  - It represents the long-term movement in time data after other components have been accounted for.
  - It tells whether a particular data set (say GDP, oil prices or stock prices) have increased or decreased over the period of time

Trend models – time series

- Passengers in one of airlines (an examples from Statistica handbook)
- Trend line = 87,65 + 2,66·t



Other Regression-Based Models

- 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

Logistic regression

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

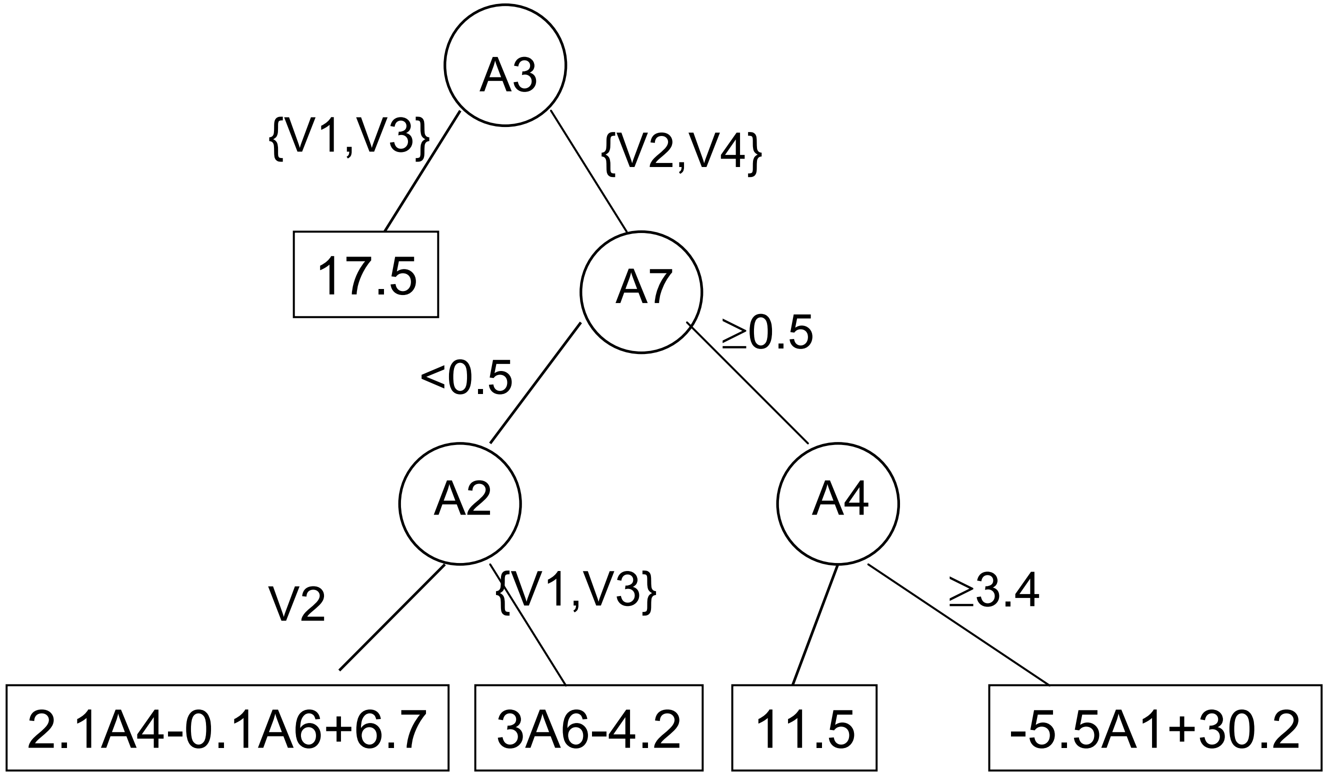
$$\log\left(\frac{P(C_1)}{1-P(C_1)}\right) = a_0 + x_1 a_1 + x_2 a_2 + \ldots + x_k a_k$$

P= Class probability

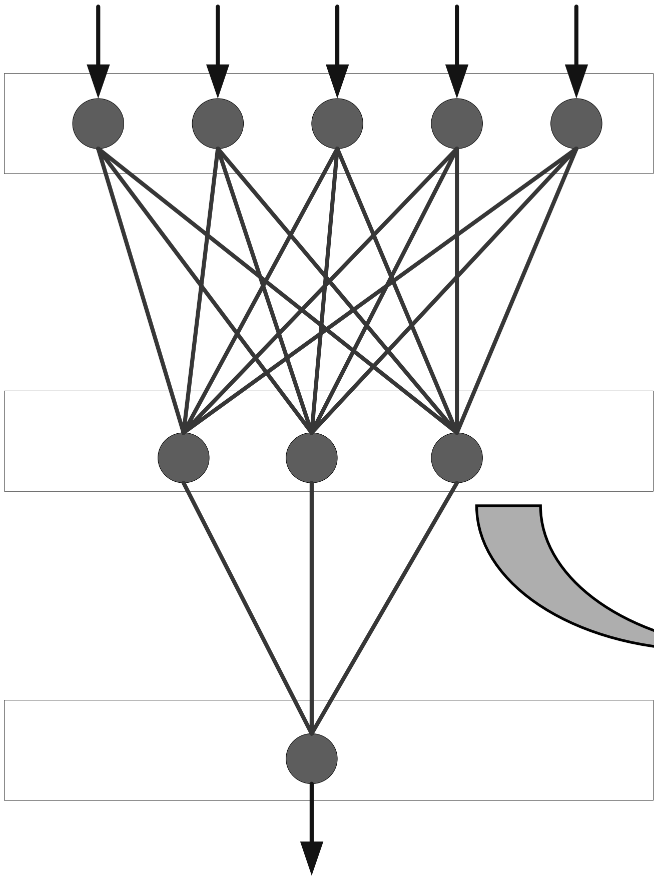
Regression Trees and Model Trees

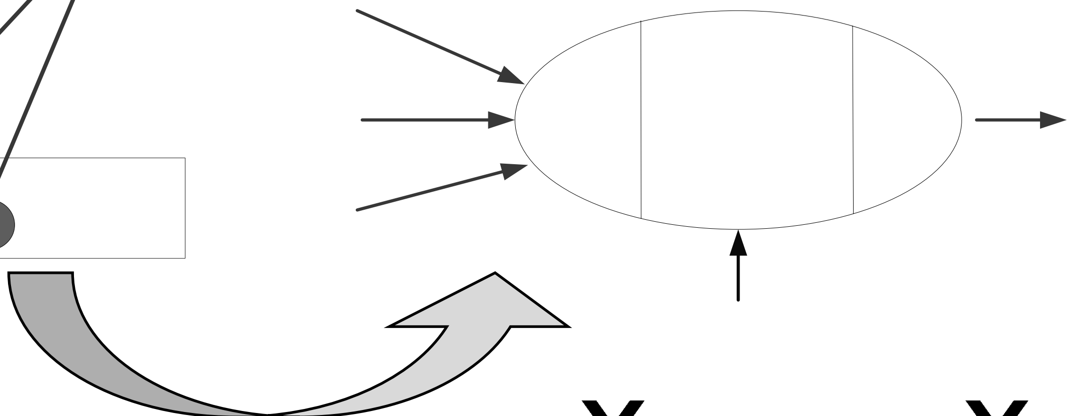
- Regression tree: proposed in **CART** system (Breiman et al. 1984)
  - CART: Classification And Regression Trees
  - Each leaf stores a *continuous-valued prediction*
  - It is the *average value of the predicted attribute* for the training tuples that reach the leaf
- Model tree: proposed by Quinlan (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



General Structure of ANN



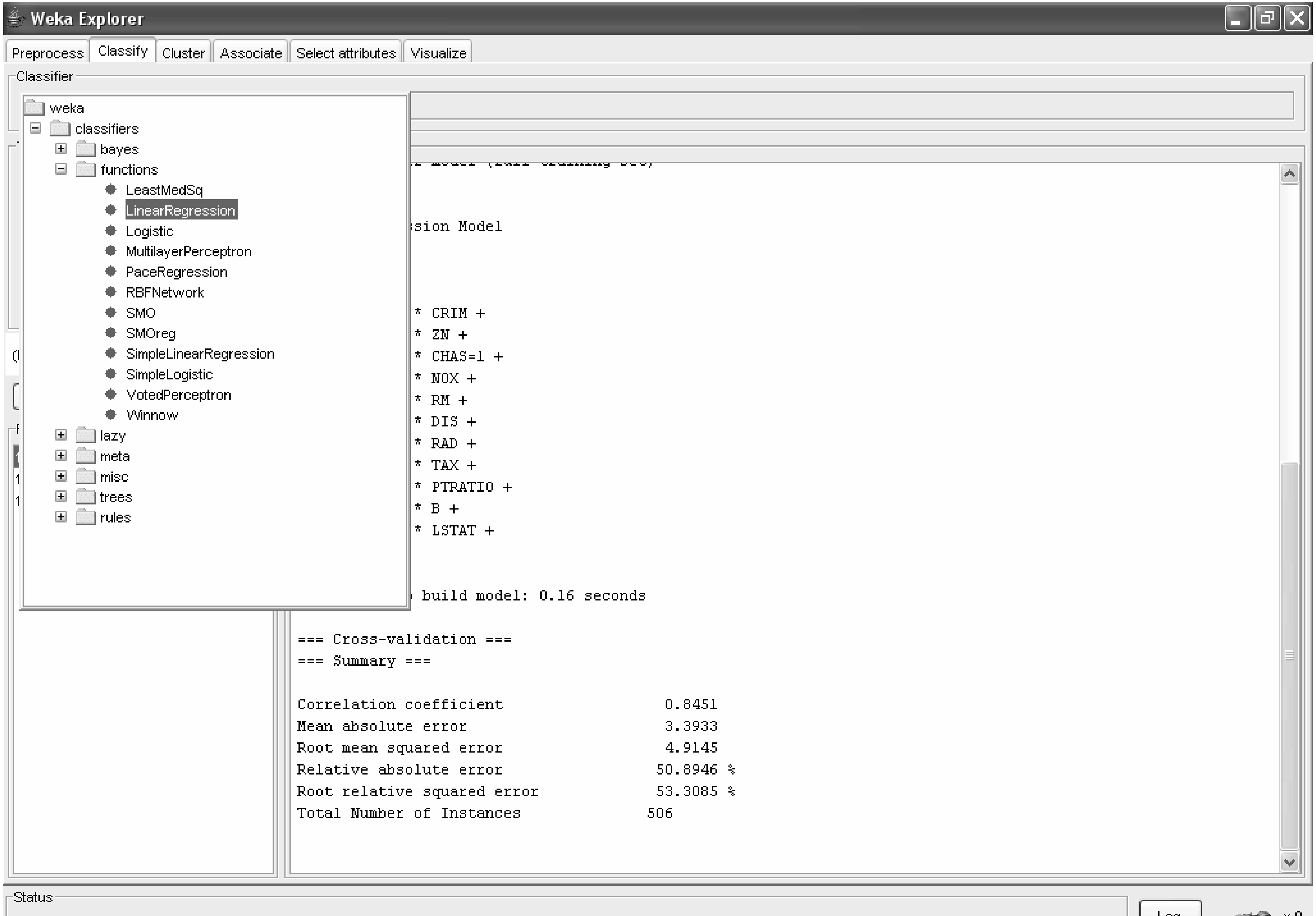


X_1 X_2

Training ANN means learning
the weights of the neurons

Input

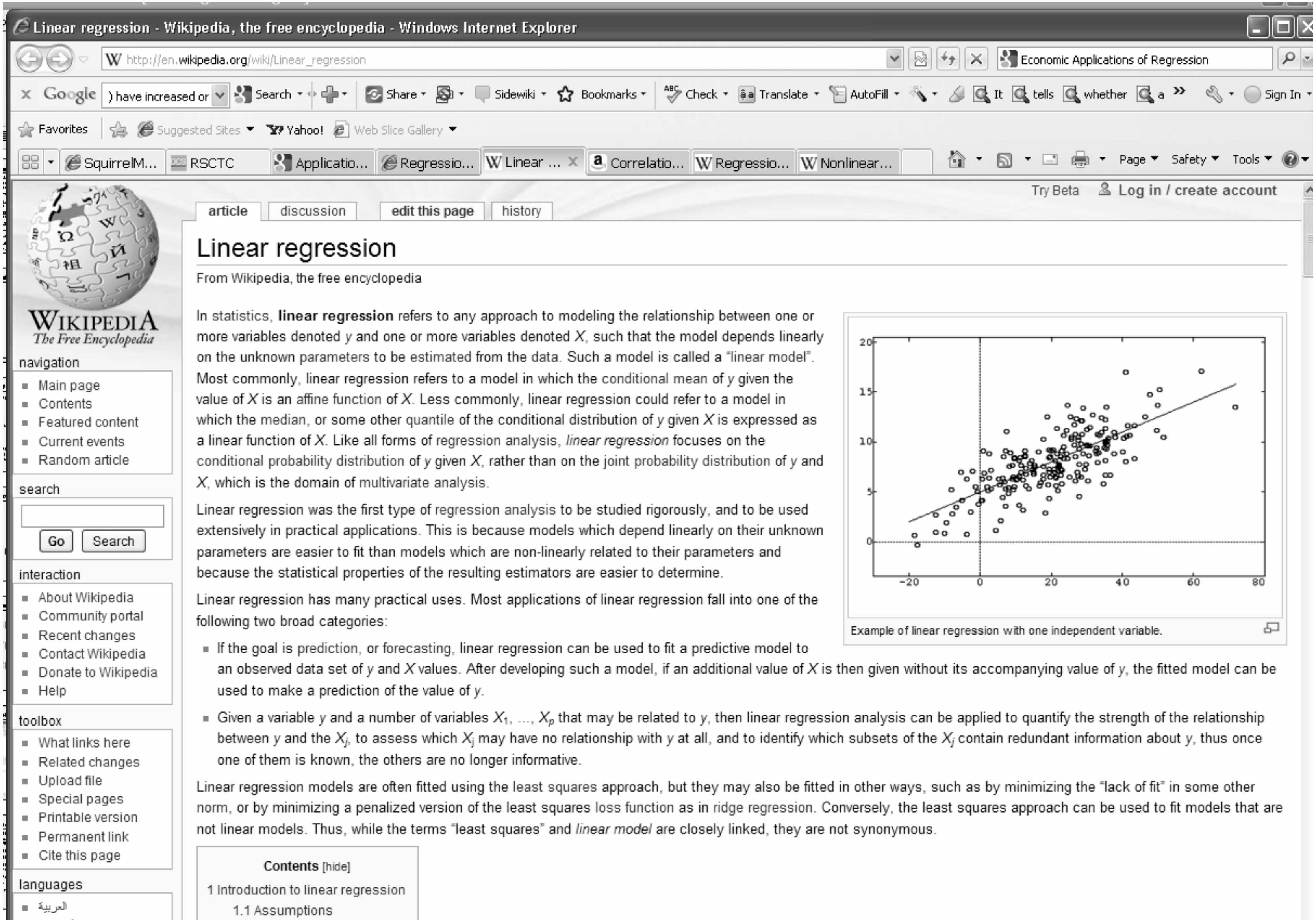
Modele predykcji zmiennej liczbowej w WEKA



Literatura / Polish Language coursebooks

- Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.
- Statystyka w zarządzaniu, A.Aczel, PWN 2000.
- Statystyka praktyczna. W.Starzyńska,
- Statystyka. Ekonometria. Prognozowanie. Ćwiczenia z Excelem. A. Snarska, Wydawnictwo Placet 2005.
- Przystępny kurs statystyki, Stanisz A., 1997.
  - Tom 2 $\rightarrow$ poświęcony wyłącznie analizie regresji!
- I wiele innych …

Verify some references from Web pages


Linear regression - Wikipedia, the free encyclopedia - Windows Internet Explorer

http://en.wikipedia.org/wiki/Linear_regression

article | discussion | edit this page | history

Try Beta | Log in / create account

Linear regression

From Wikipedia, the free encyclopedia

In statistics, **linear regression** refers to any approach to modeling the relationship between one or more variables denoted *y* and one or more variables denoted *X*, such that the model depends linearly on the unknown parameters to be estimated from the data. Such a model is called a "linear model". Most commonly, linear regression refers to a model in which the conditional mean of *y* given the value of *X* is an affine function of *X*. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of *y* given *X* is expressed as a linear function of *X*. Like all forms of regression analysis, *linear regression* focuses on the conditional probability distribution of *y* given *X*, rather than on the joint probability distribution of *y* and *X*, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
- Given a variable *y* and a number of variables $X_1, ..., X_p$ that may be related to *y*, then linear regression analysis can be applied to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*, thus once one of them is known, the others are no longer informative.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm, or by minimizing a penalized version of the least squares loss function as in ridge regression. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, while the terms "least squares" and *linear model* are closely linked, they are not synonymous.

Example of linear regression with one independent variable.

Contents [hide]