Metody predykcji – analiza regresji

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

TPD – 2008/2009

Przebieg wykładu

1. Predykcja z wykorzystaniem analizy regresji.
    1. Przypomnienie wiadomości z poprzednich przedmiotów.
2. Ocena poprawności modelu regresji liniowej.
3. Regresja wielowymiarowa.
4. Regresja nieliniowa.
5. Selekcja zmiennych.

- Uwagi: proszę odwołać się do przedmiotu „Statystyka i analiza danych” studia inżynierskie.

Modelowanie regresji

- Metoda szacowania wartości liczbowej zmiennej zależnej (objaśnianej, wynikowej) *y* na podstawie wartości zmiennych niezależnych **x**.
- Badamy zależność warunkową $y|\mathbf{x}$
- Formalnie poszukujemy modelu

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

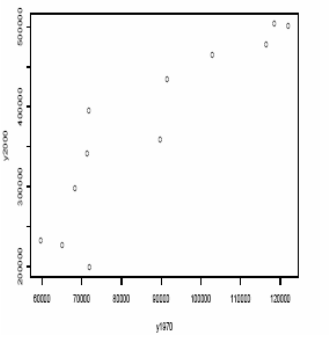
- Modele lokalne – „locally weighted regression”

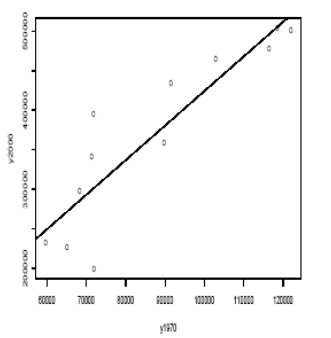
$$y = \alpha + \sum_{j=1}^{p} f_j(\mathbf{x}, \beta)$$

Przykład – ceny domów przykład z R

- W zbiorze danych *homedata* (z pakietu R) ceny 6841 domów Maplewood (New Jersey) z lat: 1970 i 2000. Interesuje nas zależność pomiędzy cenami domów z tych lat.

```
> 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
```





Regresja – model liniowy

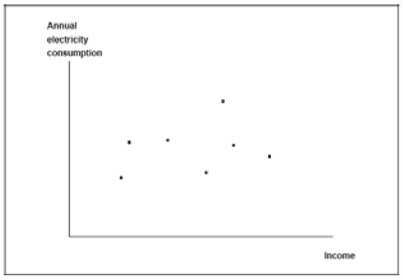
- Analityczny sposób przyporządkowania wartości zmiennej zależnej konkretnym wartościom zmiennych niezależnych.
- Liniowa regresja prosta → najprostszy rodzaj regresji, w których zależność zmiennych można opisać za pomocą linii prostej.

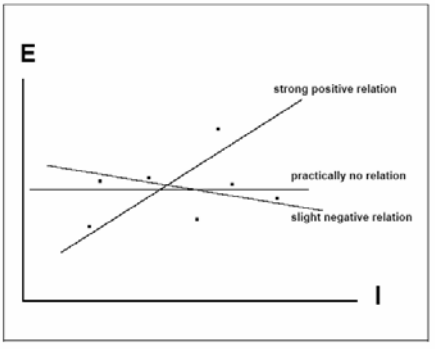
$$\hat{y} = \beta_1 \cdot x + \beta_0 + \varepsilon$$

gdzie β_1 jest *współczynnikiem kierunkowym*, β_0 wyraz wolny (punkt przecięcia z osią rzędnych); x – zmienna niezależna, y – zmienna zależna (objaśniana, przewidywana), ε - błąd losowy.

Intuicja poszukiwania regresji liniowej

- Przykład z wykładu z Ekonometrii (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.





Która linia podsumowująca ogólny trend w danych jest najlepsza?

Liniowa prosta regresji - MNK

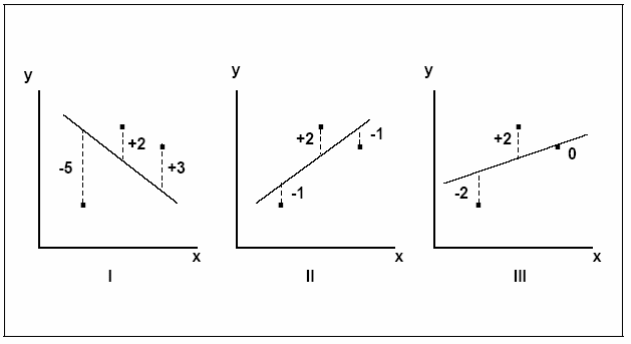
- Rzeczywiste dane $(x_1, y_1),...,(x_n, y_n).$
- Wartość teoretyczna funkcji regresji $\hat{y} = f(x)$
- Błąd oszacowania $y_i - \hat{y}_i$ tzw. wartość resztowa lub rezyduum.
- Liniowa regresja prosta → wartości rezyduów powinny być jak najmniejsze dla wszystkich i=1,…,*n*.
- Wskaźnik rozproszenia → suma kwadratów rezyduów.

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Dla liniowego wykresu dużych rezyduów nie ma być zbyt wiele → metoda najmniejszych kwadratów! daje ona najlepsze liniowe nieobciążone estymatory parametrów regresji

Przykład MNK

- Które residua (suma kwadratów) są najmniejsza?



- Proste sumowanie: I -5+2+3=0; II -1+2-1=0; III -2+2+0
- MNK: I 25+4+9=38; II: 1+4+1=6; III 4+4=8

Własności oszacowania MNK

- Linia przechodzi przez wartości średnie:

$$\hat{y} = \beta_1 \cdot \bar{x} + \beta_0 = \beta_1 \cdot \bar{x} + (\bar{y} - \beta_1 \cdot \bar{x}) = \bar{y}$$

- Wartość oczekiwana residuów jest zerowa

$$\bar{e} = \frac{\sum_{i=1}^{n} \bar{e}_i}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum y_i - \frac{1}{n}\sum(\beta_1 x + \beta_0) = \bar{y} - (\beta_1 \bar{x} + \beta_0) = \bar{y} - \bar{y} = 0$$

- Dobra własność: linia jest „średnio” właściwa.

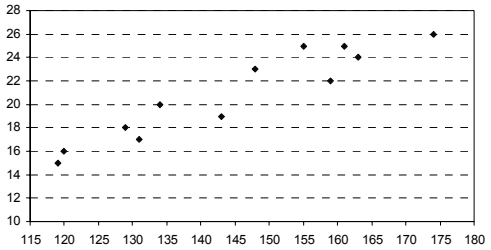
Przykład ilustracyjny (samochody)

- W firmie produkującej samochody przeprowadzono analizę sprzedaży samochodów z ostatniego miesiąca. Zebrano dane od 12 dealerów zajmujących się sprzedażą samochodów tej firmy o wielkości sprzedaży za ostatni miesiąc (zmienna zależna *Y*) oraz czasie wykupionej reklamy w ostatnim miesięcy (zmienna niezależna *X*).

Nr dealera	y	x
1	129	18
2	119	15
3	159	22
4	148	23
5	131	17
6	120	16
7	161	25
8	174	26
9	134	20
10	163	24
11	143	19
12	155	25

Samochody 2

- Wykres XY



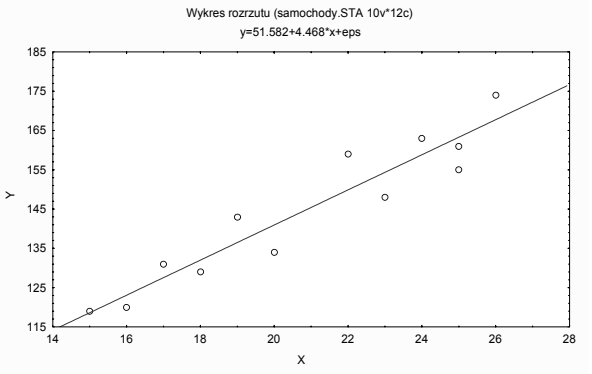
- Obliczenie współczynnika korelacji: r_{xy} = 0.9465. (statyst ist.)
- Model liniowy z oszacowanymi parametrami:

$$y = 51.584 + 4.468 \cdot x$$

- Wartość *a* oznacza, że wzrost (spadek) czasu wykupionej reklamy radiowej o jedną minutę spowoduje wzrost (spadek) sprzedaży w przybliżeniu o 4.468 sztuk samochodów.

Samochody 3

- Model $y^{\wedge} = 51.584 + 4.468 \cdot x$



Nr dealera	x	y	y^=f(x)
1	18	129	132,01
2	15	119	118,60
3	22	159	149,88
4	23	148	154,35
5	17	131	127,54
6	16	120	123,07
7	25	161	163,28
8	26	174	167,75
9	20	134	140,94
10	24	163	158,82
11	19	143	136,48
12	25	155	163,28

Równanie stochastyczne vs. deterministyczne

- Statystyczny model opisuje liczbowo zależność pomiędzy zmienną niezależną (x) oraz zmienną zależną (y)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- gdzie $\beta 0$, $\beta 1$ – nieznane parametry f.regresji, które należy oszacować; ε - składnik losowy.
- Parametry funkcji regresji nie są znane (obserwowane), podobnie jak składnik losowy, dlatego jest to równanie stochastyczne.
- Równanie deterministyczne po zastosowaniu MNK

$$\hat{y}_i = b_0 + b_1 x_i$$

- Gdzie b0, b1 oceny estymatorów parametrów funkcji regresji
  i – numer obserwacji.

Definicje zadania analizy regresji

- Wyjaśnienie w sposób analityczny kształtowania się wartości jednej zmiennej losowej (zmiennej zależnej lub objaśnianej) pod wpływam innej zmiennej (niezależnej lub objaśniającej) lub innych zmiennych.
- „Jeżeli zmienna losowa *Y* składa się z dwóch składowych: pewnej zmiennej losowej ε oraz elementu systematycznego *f*(*X*) zależnego od zmiennej *X*, to regresją zmiennej losowej *Y* względem *X* jest równanie *E*(*Y*| *X*) = *f*(*X*), przy czym zakłada się, że *E*(ε)=0”
  - Defincja [Słownik statystyczny. Kendall, Buckland]
- Regresja prosta $Y = \hat{Y} + \varepsilon$

  gdzie $\hat{Y} = f(X)$ oznacza teoretyczne poziomy zmiennej odczytane z funkcji regresji
- Funkcje – kształt liniowy lub nieliniowy

Zapis wektorowy

- Ogólna postać

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

- Rozwiązanie MNK

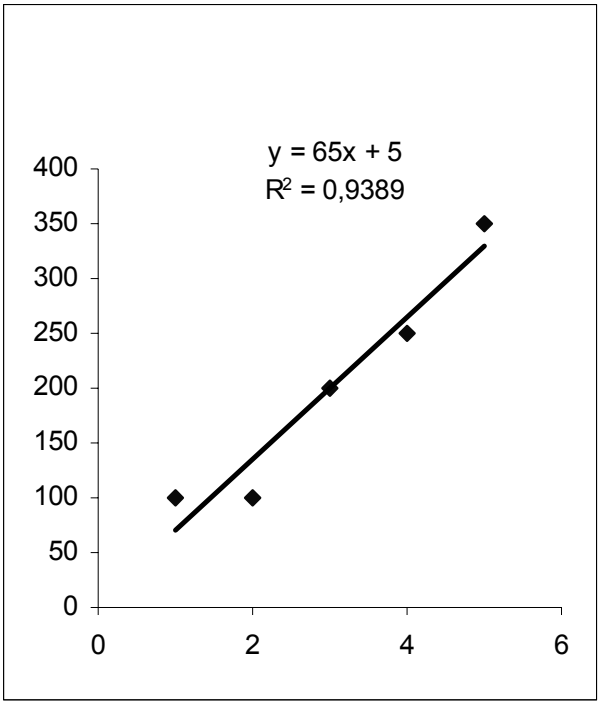
$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

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

Przykład

W celu zbadania zależności między zyskami pewnej firmy a wydatkami na szkolenia handlowców, dokonano porównania wyników dla 5 kwartałów (x_i - wydatki na szkolenia handlowców w tys. zł, y_i – zyski firmy w tys. zł):

x	1	2	3	4	5
y	100	100	200	250	350



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

$\mathbf{X^TX}=\begin{bmatrix}5&15\\15&55\end{bmatrix}$ $\mathbf{detX^TX}=50$

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

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

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

$$\mathbf{y} = \begin{bmatrix} 70 \\ 135 \\ 200 \\ 265 \\ 330 \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} 30 \\ -35 \\ 0 \\ -15 \\ 20 \end{bmatrix} \qquad \mathbf{e}^T = \begin{bmatrix} 30 & -35 & 0 & -15 & 20 \end{bmatrix}$$

$$\mathbf{e}^T\mathbf{e} = 2750$$

$$S_e^2 = 917 \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1,1 & -0,3 \\ -0,3 & 0,1 \end{bmatrix} \qquad S_e = 30,3$$

$$S(b_0) = 31,75$$

$$S_y = 9,74 \qquad S_y^2 = 95$$

$$S(b_1) = 9,58$$

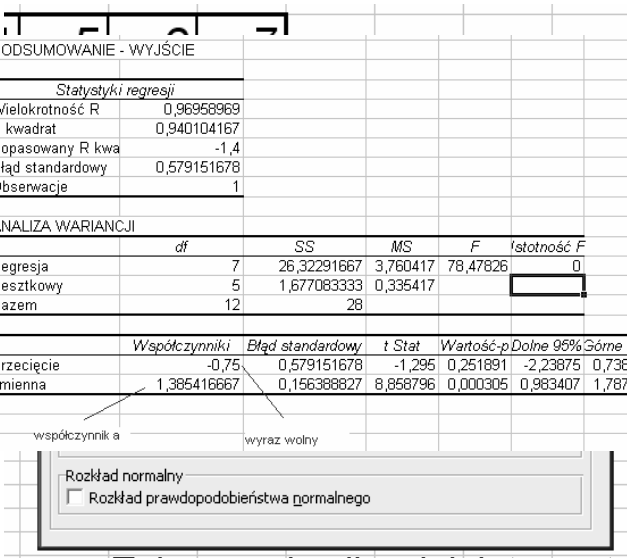
$$R^2 = 1 - \frac{2750}{5 * 9000} = 1 - 0,06 = 0,94 = 94\%$$

Co zrobimy w Excelu?
Funkcje stat. REGLINP lub dodatek Analiza Danych

10	PODSUMOWANIE - WYJŚCIE								
11									
12	*Statystyki regresji*								
13	Wielokrotność R	0,96958969							
14	R kwadrat	0,940104167							
15	Dopasowany R kwa	-1,4							
16	Błąd standardowy	0,579151678							
17	Obserwacje	1							
18									
19	ANALIZA WARIANCJI								
20		*df*	*SS*	*MS*	*F*	*Istotność F*			
21	Regresja	7	26,32291667	3,760417	78,47826	0			
22	Resztkowy	5	1,677083333	0,335417					
23	Razem	12	28						
24									
25		*Współczynniki*	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*olne 95,0%*	*órne 95,0%*
26	Przecięcie	-0,75	0,579151678	-1,295	0,251891	-2,23875	0,738754	-2,23875	0,738754
27	Zmienna	1,385416667	0,156388827	8,858796	0,000305	0,983407	1,787426	0,983407	1,787426
28									
29									
30									

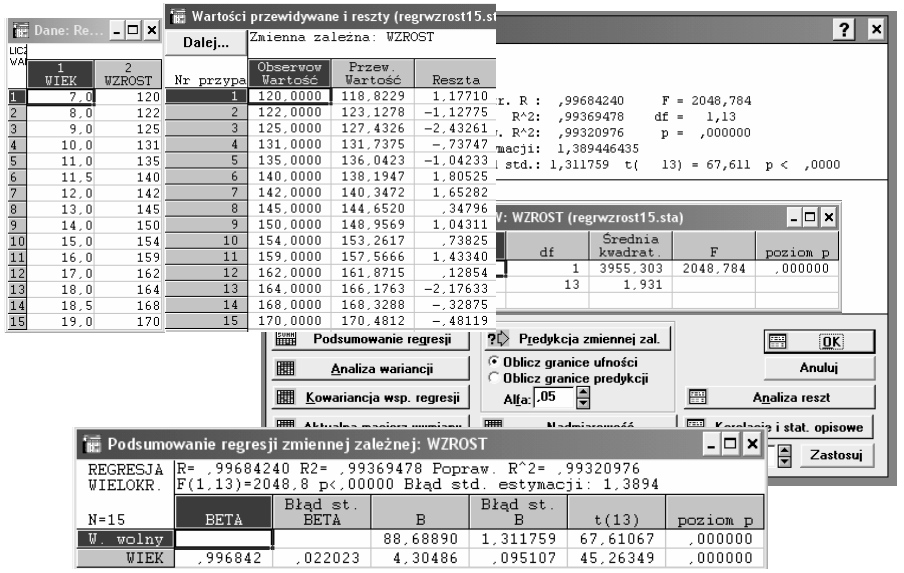
współczynnik a

wyraz wolny



Tak przy okazji → jak interpretować wyniki?

Przykład wzrost = *f*(wiek) / Statistica (Statsoft)



Weryfikacja modelu regresji

- Ocena dopasowania funkcji regresji do danych empirycznych.
- Składnik resztowy $e_i = y_i - \hat{y}_i$

  tym większy, im większy jest składnik losowy ε,

  może także wynikać z błędnego przyjęcia danej funkcji regresji.

Rozkład całkowitej zmienności zmiennej objaśnianej

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

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

Ocena modelu regresji

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

$$SST = SSR + SSE$$

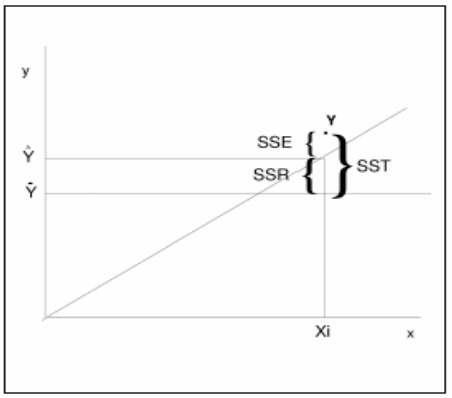
$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

gdzie

- *SSR* – regresyjna suma kwadratów odchyleń (część wyjaśniona przez zbudowany model),
- *SSE* – resztowa suma kwadratów odchyleń (część nie wyjaśniona przez zbudowany model).

Na ile dobra jest regresja?

- Współczynnik determinacji jest opisową miarą siły liniowego związku między zmiennymi, czyli miarą dopasowania linii regresji do danych.



Rysunek 2.6: Współczynnik determinacji R^2

współczynnik determinacji --- 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*.

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

Miary dopasowania modelu regresji do danych

- Współczynnik determinacji:

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

- Najważniejsza miara dopasowania funkcji regresji do danych empirycznych;
  Jest to stosunek zmienności wyjaśnianej przez model do zmienności całkowitej.
- Średni błąd kwadratowy:

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

- Wariancja resztowa (k liczba zmiennych)

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

- Błędy standardowe parametrów *b*i :

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

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

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

- odchylenie standardowe składnika resztowego – standardowy błąd oszacowania

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

Samochody 4

- R^2 = 0.8958, S = 6.1258
- R^2 ozn., że 89.58% zmienności zmiennej *y* zostało wyjaśnione przez zbudowany model.
- *S* – przeciętne odchylenie wartości empirycznych od wartości teoretycznych (wynikających ze zbudowanego modelu) wynosi 6.1258 sztuk samochodów.

Założenia modelu regresji

- Związek między *x* i *y* jest liniowy.
- Wartości zmiennej niezależnej nie są losowe. Losowość wartości *y* pochodzi wyłącznie ze składnika losowego.
- Składniki (błędy) losowe mają rozkład normalny o średniej 0 i o stałej wariancji σ^2

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

Weryfikacja – uwagi ogólne

- **Statystyczna** $\rightarrow$ dotyczy przede wszystkim weryfikacji przyjętych założeń o stochastycznej strukturze modelu oraz założeń o istotnym wpływie zmiennych objaśniających na zmienną objaśnianą za pomocą znanych testów statystycznych.
- **Merytoryczna** $\rightarrow$ wiąże się z odpowiedzią na pytanie, czy oszacowane oceny parametrów równania zgodne są z przyjętymi założeniami, a także czy istnieje możliwość "sensownej" interpretacji otrzymanych wartości ocen parametrów.

Weryfikacja modelu regresji

- Zbadaj czy istnieje związek między średnią wydajnością (mierzoną liczbą wykonanych detali określonego typu) a stażem pracy (mierzonym w miesiącach).

n	Wydaj-ność y	Staż pr. *X*
1	2	1
2	1	2
3	2	3
4	5	4
5	8	5
6	14	6
7	17	7

Załóżmy model liniowy:

$$y = \beta_0 + \beta_1 \cdot x_1 + \varepsilon$$

Wyniki obliczeń (Statistica)

Podsumowanie regresji zmiennej zależnej: WYDAJNOS

REGRESJA WIELOKR. N=7 | R= ,93930382 R^2= ,88229167 Popraw. R^2= ,85875000 F(1,5)=37,478 p<,00169 Błąd std. estymacji: 2,3770

	BETA	Błąd st. BETA	B	Błąd st. B	t(5)	poziom p
W. wolny			-4,00000	2,008909	-1,99113	,103101
STAZ_PR	,939304	,153433	2,75000	,449206	6,12192	,001687

Hipotezy dotyczące poszczególnych parametrów modelu

- Ocena poszczególnych parametrów β_i w modelu (ocena zachodzenia związku liniowego między zmienną *x* a *y*).
- Test statystyczny $H_0: \quad \beta_i = 0$

  $H_1: \quad \beta_i \neq 0$
- Statystyka testowa:

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

- Intuicja
- Badamy dla każdego parametru strukturalnego osobno, czy istotnie różni się on od zera. Jeśli nie uda nam się odrzucić hipotezy zerowej, będzie to oznaczało, że zmienna objaśniająca przy której stoi dany parametr nie wpływa na zmienną objaśnianą, więc można ją usunąć z modelu (jednakże to wymaga powtórnego oszacowania modelu, z już z aktualnym zestawem zmiennych objaśniających).

Testy istotności

2.5.3.3. Test istotności parametru modelu

Ten test dotyczy istotności pojedynczego parametru w modelu. W tym teście hipotezy formułuje się następująco:

H_0 $b_i = 0$, czyli i–ta zmienna nie ma wpływu na wynik
H_1 $b_i \neq 0$, czyli i–ta zmienna ma wpływ na wynik

Jeśli nie ma podstaw do odrzucenia hipotezy zerowej, to znaczy że nie można metodami statystycznymi uzasadnić wpływu zmiennej x_i na zmienną y. Do obliczeń definiuje się statystykę, która ma rozkład t-studenta.

2.5.3.4. Globalny test istotności

W tym teście hipotezy definiuje się następująco:
H_0 $b_0 = b_1 = \ldots = b_{m-1} = 0$
H_1 Przynajmniej jeden $b_i \neq 0$
Do obliczeń korzysta się ze statystyki testowej F o rozkładzie F-Snedocera o $(n - m - 1)$ stopniach swobody:

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

Istotność modelu regresji dla przykładu samochodowego.

- Model $y = 51.584 + 4.468 \cdot x$

Źródło zmienności	Liczba stopni swobody	Suma kwadratów odchyleń	Przeciętna suma kwadratów odchyleń
Model (część wyjaśniona)	(k=1) 1	*SSR* 3227.4151	(*MSR*=*SSR*/1) 3227.4151
Błąd (część niewyjaśniona)	($n – k – 1 = n$-2) 10	*SSE* 375.2515	(*MSE*=*SSE*/(n-2)) 37.5252
Całkowita	(n-1) 11	*SST* 3602.67	

- $R2 = 0.8958$, $S = 6.1258$, $F = 86.0067$
- Wartość krytyczna statystyki z tablic rozkładu F przy poziomie istotności $\alpha = 0.05$ wynosi 4.96
- Podsumujmy wyniki:
  - Model jest statystycznie istotny.

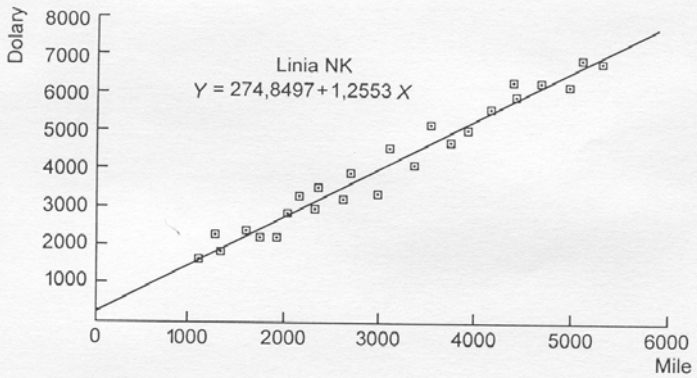
Przykład American Express

- Rozważmy przykład posiadaczy kart kredytowych American Express $\rightarrow$ 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 $)
1211	1802
1345	2405
1422	2005
1687	2511
1849	2332
2026	2305
2133	3016
2253	3385
2400	3090
2468	3694
2699	3371
2806	3998
3082	3555
3209	4692
3466	4244
3643	5298
3852	4801
4033	5147
4267	5738
4498	6420
4533	6059
4804	6426
5090	6321
5233	7026
5439	6964

Analiza regresji – American Express



Rysunek 10.12. Linia NK w badaniu zleconym przez American Express

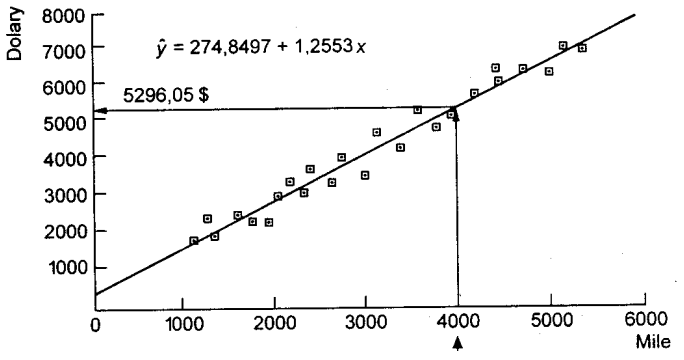
Weryfikacja równania regresji

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

  $S(b_1)$ = 0.00497
- Współczynnik determinacji R^2 = 0.9652

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:

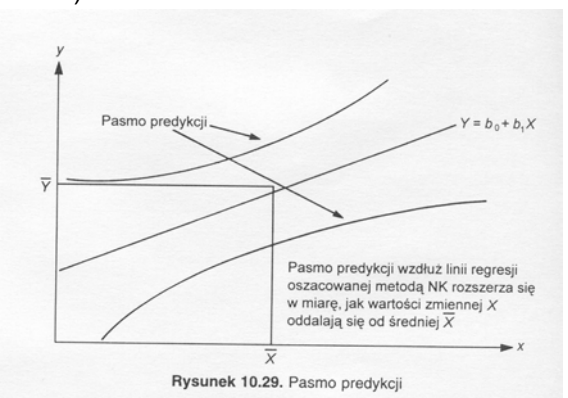
  $\bar{x}$ = 79448/25=3177,92; SS_x = 40947557,84 a *s* = 318,16
  Ponadto *t* przy 23 stopniach swobody wynosi 2,069

Stąd przedział 5296,05±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$.

Przedziały predykcji

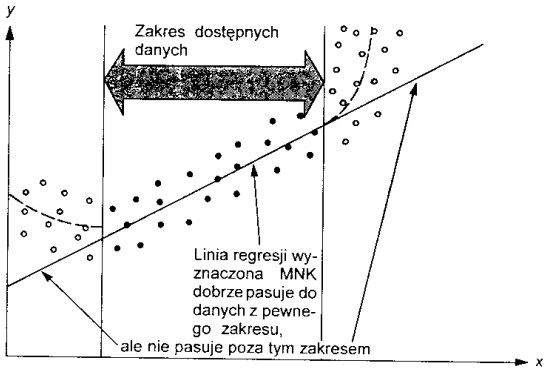
- Ograniczenie prognoz punktowych → błędu pochodzące zarówno z niepewności szacunków, jak i losowej zmienności położenia punktów w stosunku do linii regresji.
- Stosuj wtedy tzw. przedziały predykcji (tzw. prognozy przedziałowe).



Rysunek 10.29. Pasmo predykcji

Przewidywanie w regresji

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

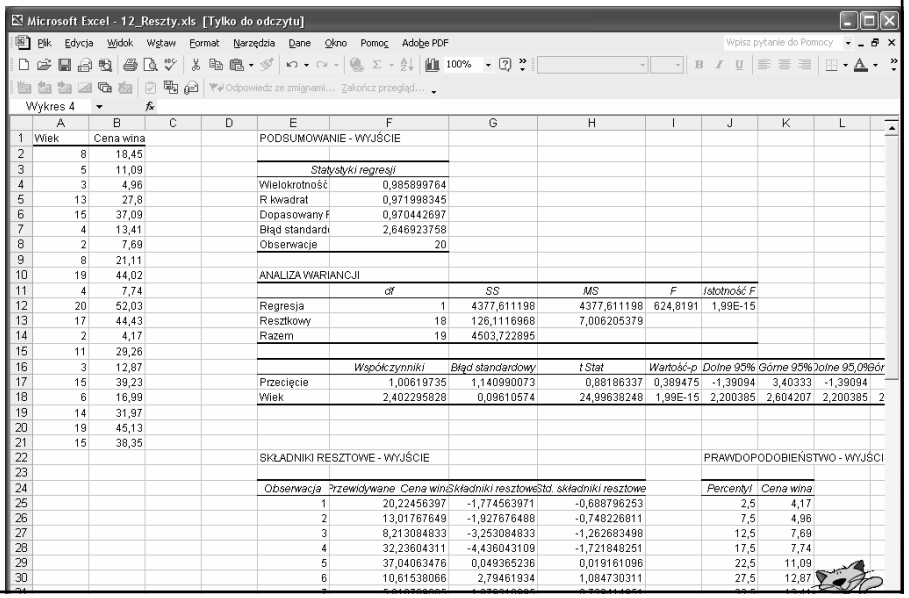


Rysunek 10.27. Niebezpieczeństwo ekstrapolacji

Rozkłady reszt

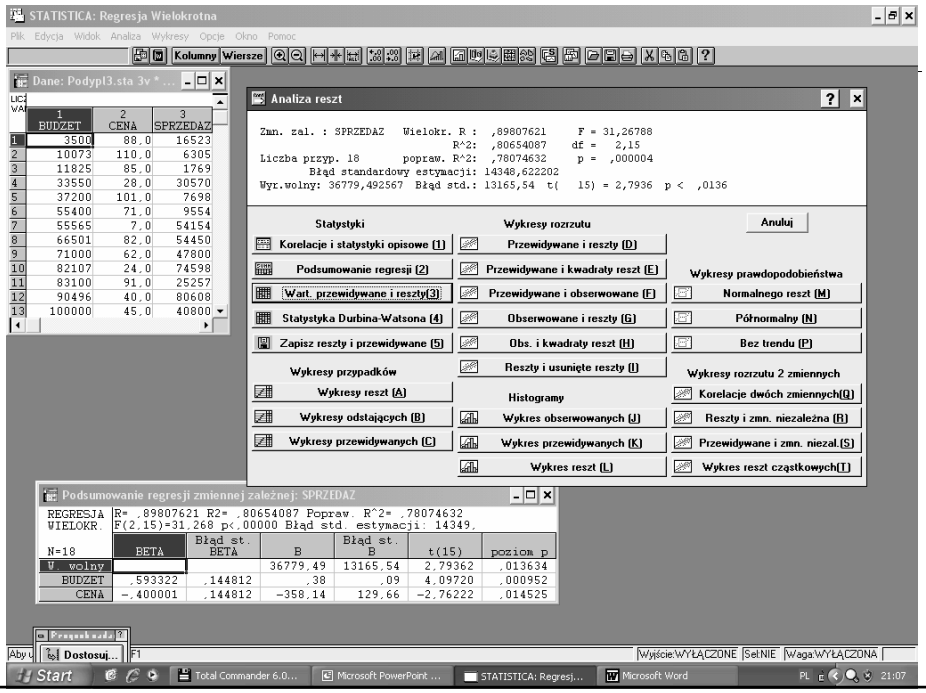
- Sposób szybkiej oceny (jakość reszt).
- Założenia modelu liniowego: Składniki (błędy) losowe mają rozkład normalny o średniej 0 i o stałej wariancji $\rightarrow$ czyli reszty powinny mieć charakterystyczny rozrzut; najlepiej obserwować to na wykresach rozrzutu reszt.

Wykresy rozkładu reszt (przykład zależności cen wina od wieku wina) = dane za A.Snarska: Statystyka, ekonometria, prognozowanie.



Microsoft Excel - 12_Reszty.xls [Tylko do odczytu]

	A	B	C	D	E	F	G	H	I	J	K	L
1	Wiek	Cena wina			PODSUMOWANIE - WYJŚCIE							
2	8	18,45										
3	5	11,09			*Statystyki regresji*							
4	3	4,96			Wielokrotność	0,985899764						
5	13	27,8			R kwadrat	0,971998345						
6	15	37,09			Dopasowany R	0,970442697						
7	4	13,41			Błąd standardo	2,646923758						
8	2	7,69			Obserwacje	20						
9	8	21,11										
10	19	44,02			ANALIZA WARIANCJI							
11	4	7,74				*df*	*SS*	*MS*	*F*	*Istotność F*		
12	20	52,03			Regresja	1	4377,611198	4377,611198	624,8191	1,99E-15		
13	17	44,43			Resztkowy	18	126,1116968	7,006205379				
14	2	4,17			Razem	19	4503,722895					
15	11	29,26										
16	3	12,87				*Współczynniki*	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95,0%*	*Gór*
17	15	39,23			Przecięcie	1,00619735	1,140990073	0,88186337	0,389475	-1,39094	3,40333	-1,39094	
18	6	16,99			Wiek	2,402295828	0,09610574	24,99638248	1,99E-15	2,200385	2,604207	2,200385	2
19	14	31,97											
20	19	45,13											
21	15	38,35											
22					SKŁADNIKI RESZTOWE - WYJŚCIE					PRAWDOPODOBIEŃSTWO - WYJŚCI			
23													
24					*Obserwacja*	*Przewidywane Cena wina*	*Składniki resztowe*	*Std. składniki resztowe*		*Percentyl*	*Cena wina*		
25					1	20,22456397	-1,774563971	-0,688796253		2,5	4,17		
26					2	13,01767649	-1,927676488	-0,748226811		7,5	4,96		
27					3	8,213084833	-3,253084833	-1,262683498		12,5	7,69		
28					4	32,23604311	-4,436043109	-1,721848251		17,5	7,74		
29					5	37,04063476	0,049365236	0,019161096		22,5	11,09		
30					6	10,61538066	2,79461934	1,084730311		27,5	12,87		



STATISTICA: Regresja Wielokrotna

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,

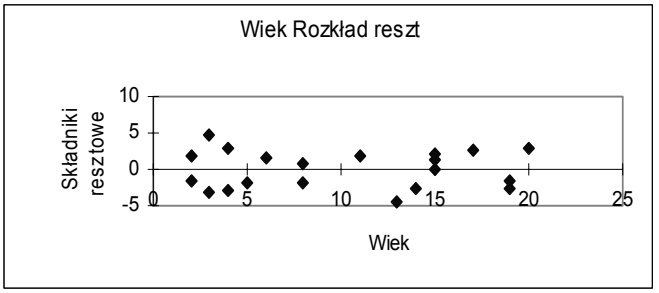
N=18	BETA	Błąd st. BETA	B	Błąd st. B	t(15)	poziom p
U wolny			36779,49	13165,54	2,79362	,013634
BUDZET	,593322	,144812	,38	,09	4,09720	,000952
CENA	-,400001	,144812	-358,14	129,66	-2,76222	,014525

Dostosuj...

Start | Total Commander 6.0... | Microsoft PowerPoint ... | STATISTICA: Regresj... | Microsoft Word | PL | 21:07

Wykres rozkładu reszt

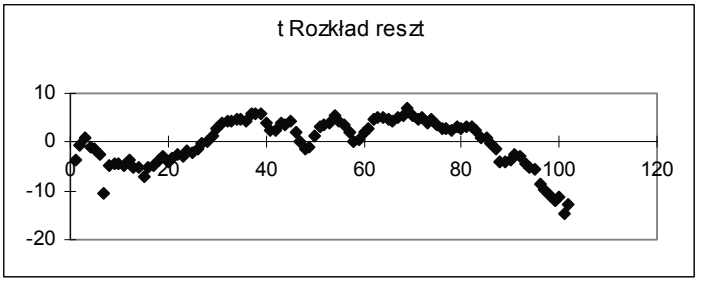
- Wina / Składniki resztowe w zależności od wieku



- Reszty przypuszczalnie spełniają założenia modelu regresji. Rozproszenie nieregularne ale w „pasie” o pewnej szerokości. Brak korelacji wzajemnej kolejnych składników.

Wykres rozkładu reszt – zestaw 2

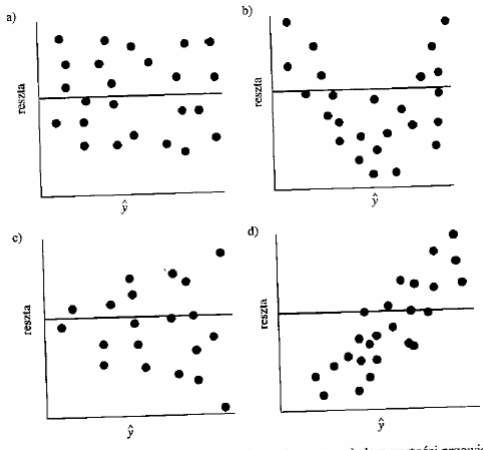
- Inny przykład wykresu składników resztowych.



- Układ linii wykresu wskazuje, że reszty następne zależą od poprzednich i rozbiegają się poza „ograniczony pas”.

Wykresy reszt – różne interpretacje

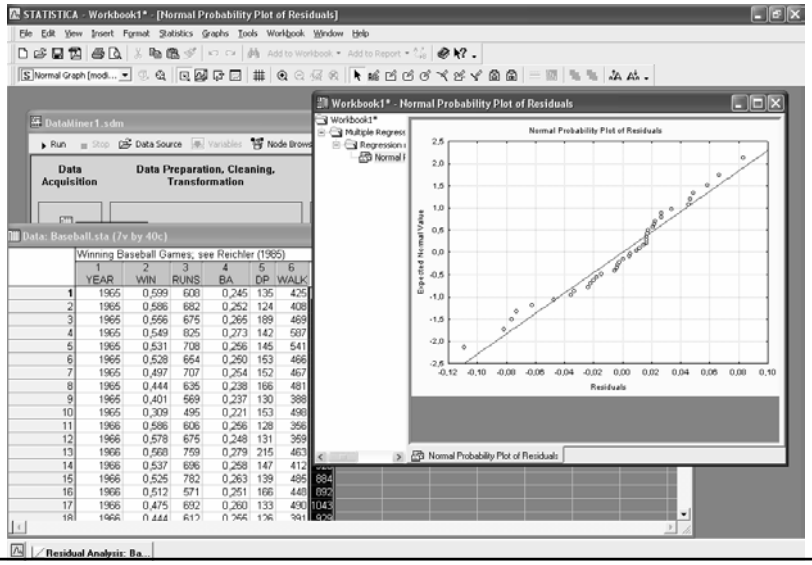
- Oceń poniższe sytuacje



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

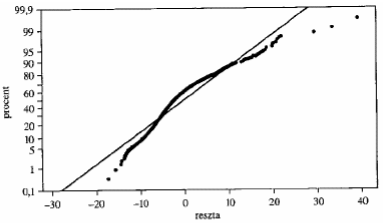
Sprawdzenie wykresu kwantylowego

- Dataminer 7 (Normality Probability Plot of Residuals)

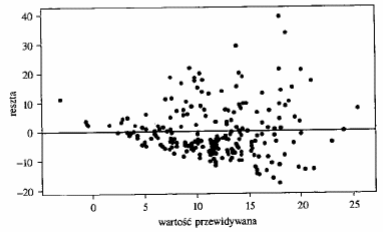


Inny przykład – inny „baseball American League 2002”

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

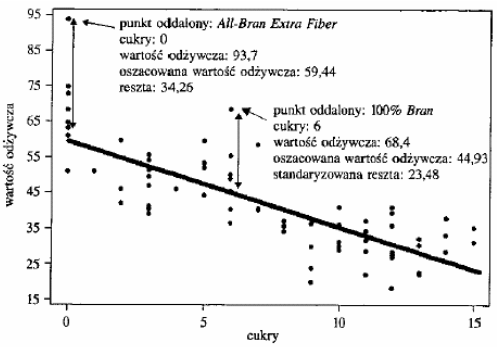


Rys. 2.16. Wykres standaryzowanych reszt względem wartości przewidywanych — naruszone założenie o stałej wariancji

- Naruszone założenia

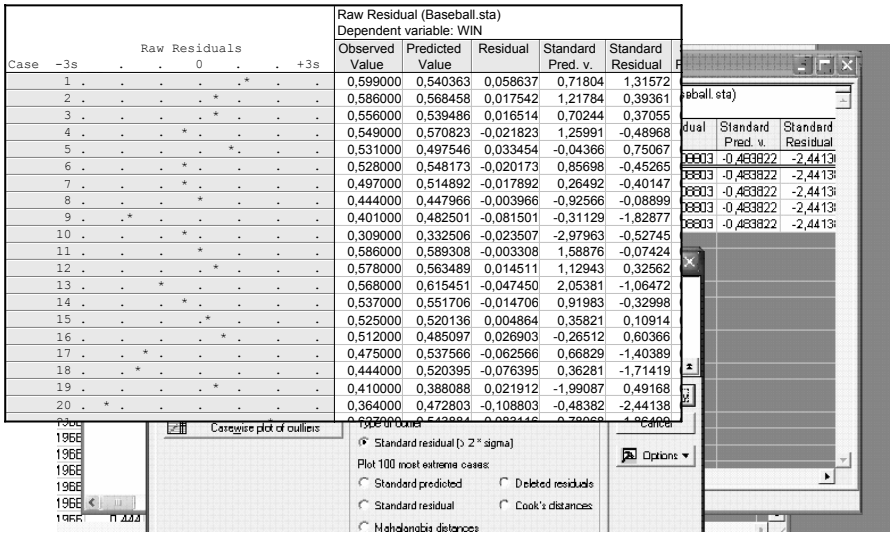
Punkty oddalone - outliers

- Przykład „płatki śniadaniowe” [Larose 08] – dwie obserwacje są zdecydowanie bardziej odlegle od linii regresji niż pozostałe → analiza reszt



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

Punkty oddalone (reszty standaryzowane)



Raw Residuals

Case	-3s . . 0 . . +3s	Raw Residual (Baseball.sta) Dependent variable: WIN 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 wielokrotna (wielowymiarowa, wieloraka)

- Zmienna objaśniana zależy od więcej niż jednej zmiennej (sytuacja częsta w praktyce).
- Model regresji zmiennej *y* względem zbioru *m-1* zmiennych niezależnych $x_1, x_2, \ldots, x_{m-1}$ jest określony równaniem:

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

- Analiza wielowymiarowa

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

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

Analiza wielowymiarowa

- Wybrane wskaźniki

$$\bar{x} = \left[\bar{x}_1 \quad \bar{x}_2 \quad \ldots \quad \bar{x}_m\right]$$

- Miara rozproszenia – macierz kowiarancji

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

Model liniowy regresji wielokrotnej

- 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

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

Regresja wielokrotn

- Dane są informacje o budżecie reklamowym pewnego produktu, jego cena jednostkowa oraz finalna sprzedaż jednostkowa.

	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 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

Korelacje (podypl3.sta)

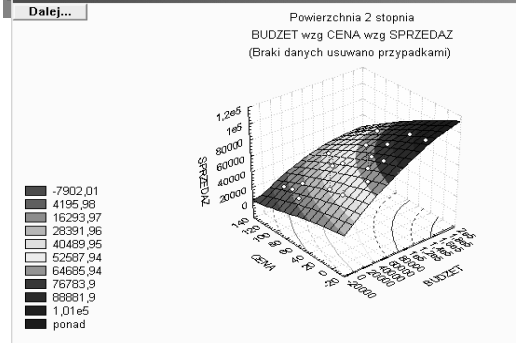
Dalej...

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

Wykres1: Powierzchnia 2 stopnia

Dalej...



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 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.

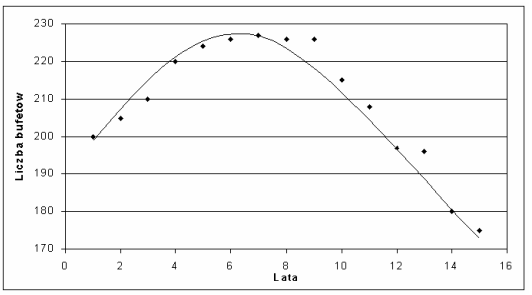
$$Y = \sum_{i=1}^{k} b_k z_k$$

$$Z_k = h_k(X)$$

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

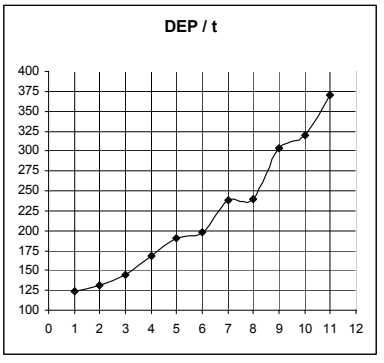
- Zakładamy, że kształt równania jest $y = a_0 + a_1 \cdot t + a_2 \cdot t^2$
- Wprowadzamy zmienne zastępcze $z_1 = t \quad z_2 = t^2$
- Rozwiązanie
  - a0=188
  - a1=11,031
  - a2=-0,814
- Weryfikacja
  - R2=0.996 s=3,37
  - Obie wartości statystyk *t < 0.05*

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

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

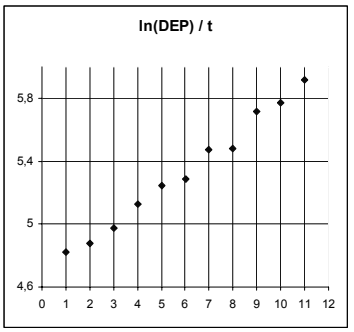


Hipoteza – wykładniczy przebieg $DEP = a \cdot e^{b \cdot t}$

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·*t*, R2=0.989, współczynniki istotne.
- Przekształcenie odwrotne

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

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

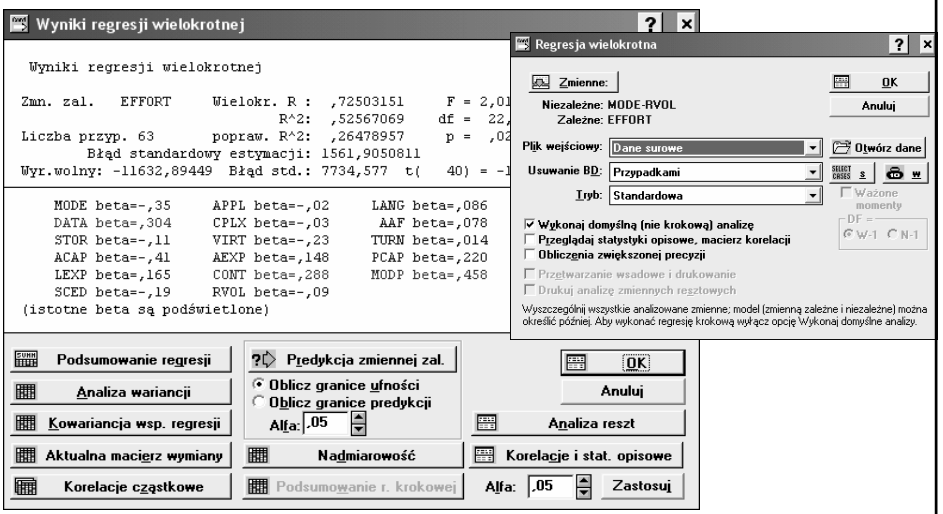
  r(y,x1)=r1=0.950 $\rightarrow$ wybieramy x1

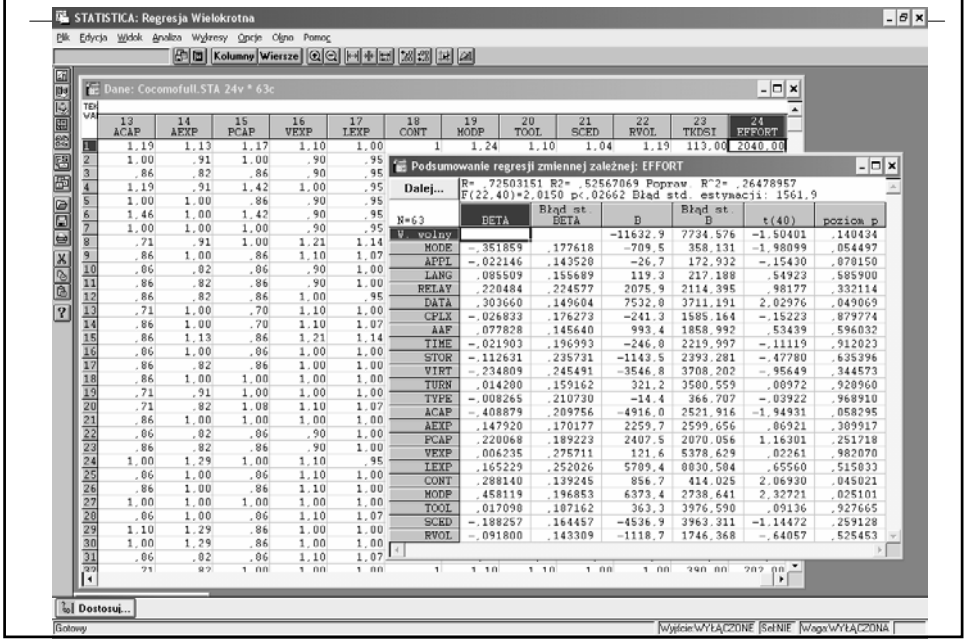
Co z pozostałymi zmiennymi?

Regresja krokowa

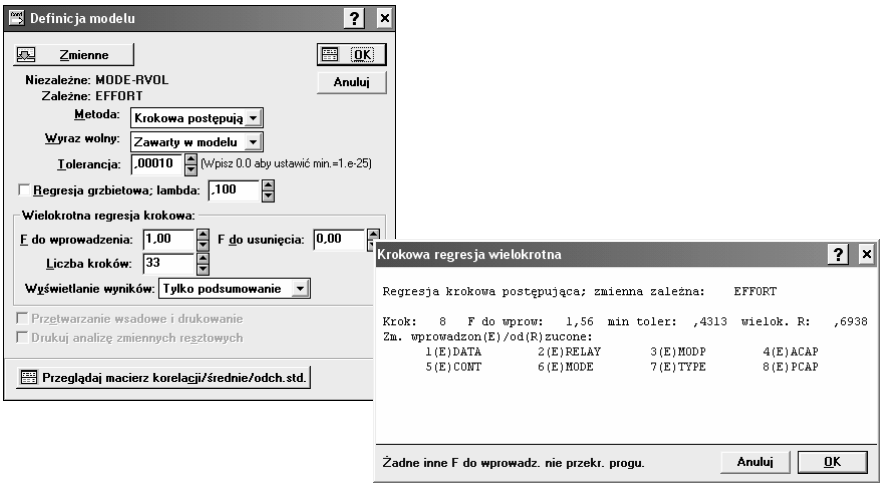
- Postępująca (*forward*)
  - Zakłada kolejne dołączanie do listy zmiennych objaśniających tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.
- Wsteczna (*backward*)
  - Usuwamy ze zbioru zmiennych, ta które mają najmniejszy wpływ na zmienną zależną.
  - Stosując R2 lub testy istotności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

Regresja wielokrotna - Statistica





Regresja krokowa



Literatura

- 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 …