Regresja wielokrotna

Model dla zależności liniowej:

$Y=a+b_1X_1+b_2X_2+...+b_nX_n$

Cząstkowe współczynniki regresji wielokrotnej: $b_{1,\dots,}\ b_n$

Zmienne niezależne (przyczynowe): $X_{1,\dots,}\ X_n$

Zmienna zależna (skutkowa): Y

i-ty, cząstkowy współczynnik regresji opisuje o ile średnio zmieni się wartość zmiennej Y przy wzroście wartości zmiennej X_i o jednostkę **przy ustalonych wartościach pozostałych zmiennych** niezależnych.

Współczynnik determinacji (*R-Square*) – informacja o tym, w jakim stopniu równanie regresji wyjaśnia zmienność zmiennej zależnej. Przyjmuje wartość od 0 do 100%. Im więcej cech zostało umieszczonych w modelu tym wyższe wartości on przyjmuje.

Poprawiony współczynnik determinacji (*adjusted R-square*) – zawiera poprawkę na liczbę zmiennych w modelu. Jeżeli dodanie zmiennej do modelu nie poprawia jakości wnioskowania, poprawiony współczynnik determinacji może być mniejszy.

Mamy m cech, więc pełny model wyglądałby:

$Y = a + b_1X_1 + b_2X_2 + ... + b_mX_m$

Można postawić Hipotezę zerową, że wszystkie współczynniki cząstkowe są równe 0 przy alternatywnej, że przynajmniej jeden nie jest.

Jednak nawet po odrzuceniu hipotezy o nieistotności modelu nie wszystkie zmienne przyczynowe ($X_1, \ldots, X_n$) wpływają (w przybliżeniu liniowo) na zmienną skutkową (Y). Działaniem statystycznym jest wybór tych zmiennych przyczynowych, które liniowo wpływają na Y.

Są różne kryteria wyboru zmiennych przyczynowych występujących w modelu. np:

- AIC (Akaike's Information Criterion) AIC = n*ln(SSE/n)+2p
- SBC (Schwarz's Bayesian Criterion) SBC = n*ln(SSE/n)+(p)*ln(n)

gdzie n jest liczbą obserwacji; p – liczbą parametrów, tj. liczbą cech + 1; SSE – sumą kwadratów odchyleń dla błędu w wybranym modelu.

Ani AIC ani SBC nie pokazują bezpośrednio, które zmienne powinny być zawarte w modelu a których tam być nie powinno. Oczywiście można sprawdzić wszystkie kombinacje (każdy podzbiór cech), tzn. policzyć wybrane kryterium (np. AIC) i wybrać podzbiór z najniższą wartością (AIC). Jednak ilość takich kombinacji jest spora (2^n, więc przy dziesięciu cechach jest 1024 kombinacje, przy 20 – ponad milion). Dlatego stosuje się metody, które choć **nie dają**

gwarancji znalezienia **najlepszego** układu cech, to szybko wskażą **wysoko oceniany** układ.

Często stosowane są **metody krokowe** – mając dany układ cech dodajemy lub usuwamy jedną cechę, tj. dodajemy cechę nie występującą obecnie w modelu którą w danym momencie uważamy za właściwą, lub usuwamy cechę występującą w modelu, jeżeli uznamy ją w danym momencie za niewskazaną.

FORWARD SELECTION Jest to metoda, która polega na stopniowym dołączaniu do modelu kolejnych zmiennych. W pierwszym kroku tworzony jest model bez zmiennych przyczynowych. W drugim – z jedną zmienną niezależną, tą, którą charakteryzuje najniższy rzeczywisty poziom istotności z nią związany (P_{value} dla hipotezy, że ta zmienna nie wyjaśnia liniowo błędów modelu). W następnym kroku tworzony jest na tej samej zasadzie model z dwiema zmiennymi niezależnymi itd.

Postępowanie trwa tak długo, aż nie zostanie znaleziona już zmienna, dla której rzeczywisty poziom istotności jest mniejszy niż zakładany (np 50%).

BACKWARD SELECTION Jest to metoda, która polega na stopniowym usuwaniu z modelu kolejnych zmiennych. W pierwszym kroku tworzony jest model z wszystkim deklarowanymi zmiennymi. Kolejne kroki polegają na usuwaniu po jednej zmiennej, która najmniej wnosi do modelu, tzn. P_{value} jest największe. Analiza trwa do momentu, gdy pozostałe w modelu zmienne charakteryzują się P_{value} poniżej zakładanego poziomu (np. 10%).

STEPWISE to połączenie powyższych metod. Określa się poziom istotności, przy którym zmienna jest dołączana bądź usuwana z modelu.

Liczba obserwacji musi być większa od liczby parametrów.

Reszty modelu (różnica między rzeczywistą a oszacowaną modelem wartością zmiennej zależnej) powinny spełniać kryteria:

- reszty posiadały rozkład normalny w każdym punkcie szacowanej (wyliczonej) wartości zmiennej zależnej
- wartość oczekiwana reszt dla każdej oszacowanej wartości (wyliczonego Y) wynosiła 0
- równa wariancja reszt dla wszystkich oszacowanych wartości zmiennej zależnej Y.

Do oceny wizualnej tych kryteriów może służyć wykres z wyliczonymi wartościami Y (*predicted value*) na osi X i resztami na osi Y.

Jeżeli niektóre cechy niezależne (X_1 – X_n) są liniowo współzależne, to mogą wyjaśniać tą samą część zmiennej niezależnej Y. Jeśli tak jest, to niektóre zmienne można usunąć z modelu jako nadmiarowe. Do wykrywania ich

służą odpowiednie statystyki (*Variance Inflation Factor* czy *Condition index*)

Przykład:

Czy zmienna Y zależy od X_1, X_2, X_3 w sposób liniowy?

X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
4	3	3	9	4	5	4	8
5	4	3	11	2	3	4	3
6	6	5	11	4	6	5	6
3	2	4	6	6	4	1	15

$Y= a+b_1X_1+b_2X_2+b_3X_3$

metoda: usuwanie po jednej zmiennej w regresji wielokrotnej.

Krok 0: wszystkie zmienne w modelu (R^2=99,5%)

Krok 1: P_{value} dla X_2 to 12%, więc usuwamy tą zmienną z modelu. Pozostają X_1 i X_3 (R^2=99%)

Krok 2: P_{value} dla X_3 to 0,5%, więc nie usuwamy tej zmiennej. Koniec selekcji zmiennych.

$Y= 3,41 + 2,16*X_1 – 1,09*X_3$

Regresja wielomianowa (krzywoliniowa)

Model dla regresji wielokrotnej liniowej:

$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n$

Jeżeli mamy cechę X, to możemy określić:

$X_1 = X$; $X_2 = X^2$; $X_3 = X^3$ itd.

Ponieważ wielomiany bazowe nie są liniowo niezależne (np. X jest liniowo współzależne z X^3) to dodanie bądź usunięcie jakiejś „cechy" ($X_i = X^i$) z modelu powodować będzie zmiany w rzeczywistym poziomie istotności dla oceny przydatności pozostałych „cech" tego modelu.

Założenia stosowalności są jak w regresji wielokrotnej, czyli np.: stopień wielomianu musi być znacznie niższy niż liczba obserwacji.

Przykład:

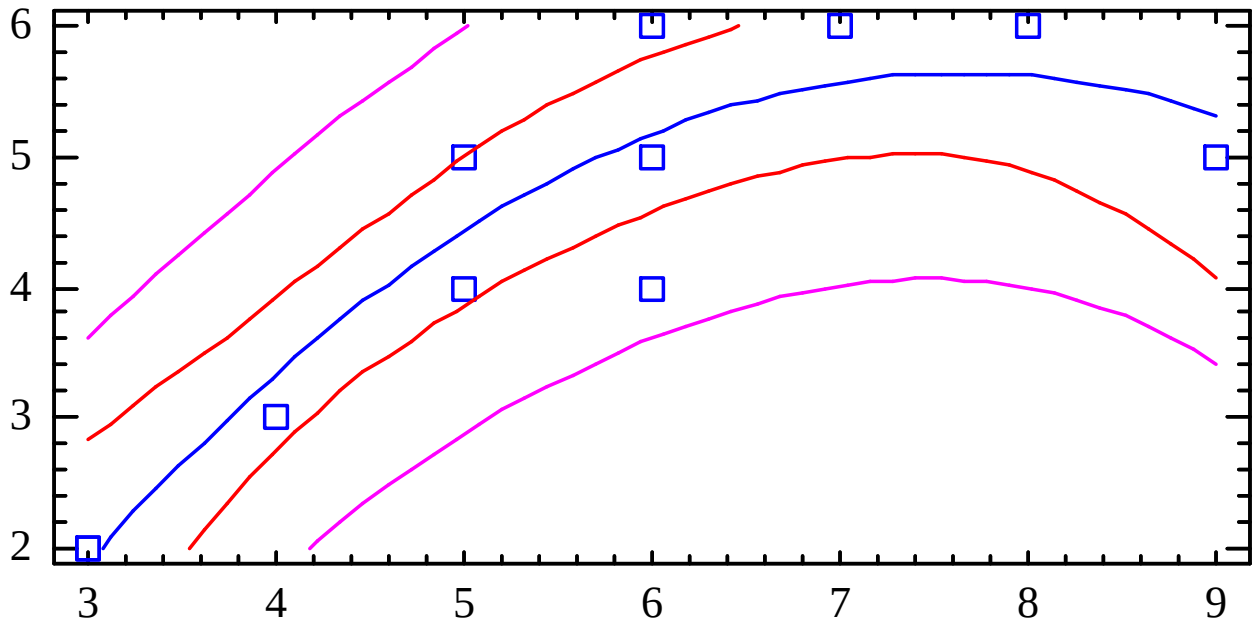
X	3	4	3	5	6
Y	2	3	2	4	4
5	6	7	6	8	9
5	5	6	6	6	5

Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	-4,5669	1,72144	-2,65295	0,0291
X	2,67406	0,623519	4,28866	0,0027
X ^2	-0,175016	0,0529431	-3,30573	0,0108

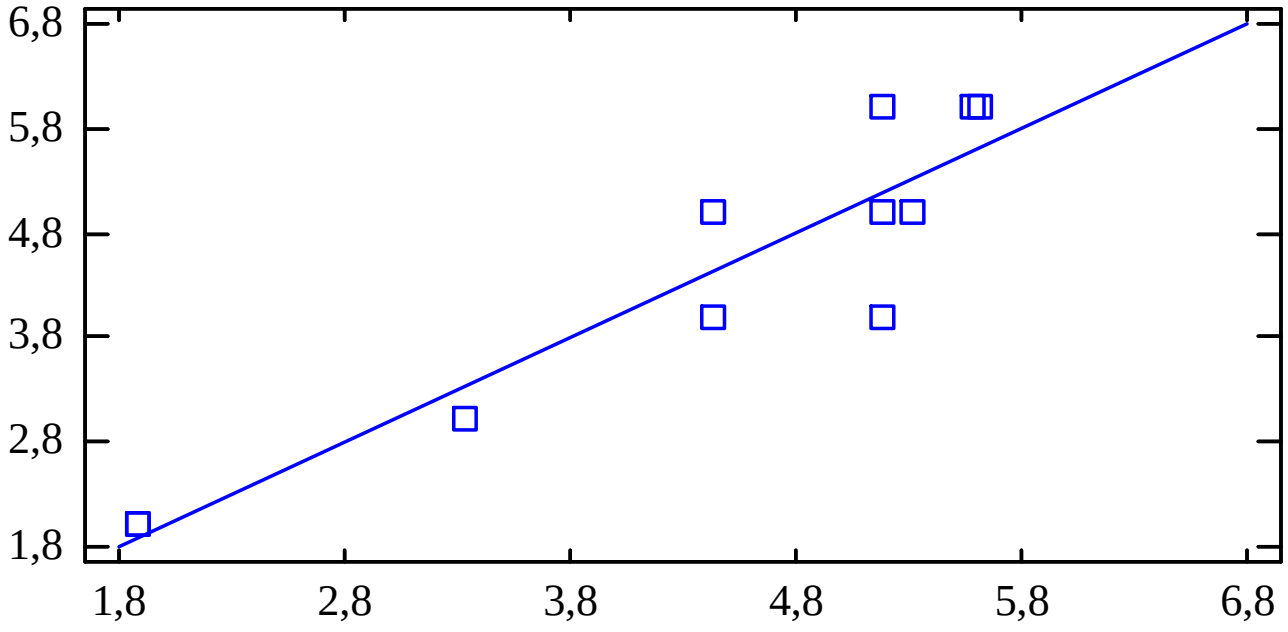
Przyjmując jako model wielomian stopnia 2 utyskujemy powyższą tabelę wariancji dla poszczególnych składników.

Y = –4,5669 + 2,67406*X – 0,175016*X^2

Wykres przedstawia zależność zmiennej zależnej Y (pionowa oś) od zmiennej niezależnej X (pozioma oś)



Wykres pokazuje jak różnią się wartości obserwowane (oś Y) od wartości przewidywanych modelem (oś X)



Regresja pojedyncza nieliniowa

Zależność Y od X nie jest liniowa. Znamy (podejrzewamy) charakter tej zależności (kwadratowa, logarytmiczna, ...) — wzór funkcji f i g.

$E(f(Y)|X)=a+b*g(X)$

zamiast $E(Y|X)=a+bX$ jak w regresji prostej

Po transformacji zmiennej skutkowej Y i zmiennej przyczynowej X funkcjami f i g estymacja parametrów a i b następuje tak jak w regresji prostej, np:

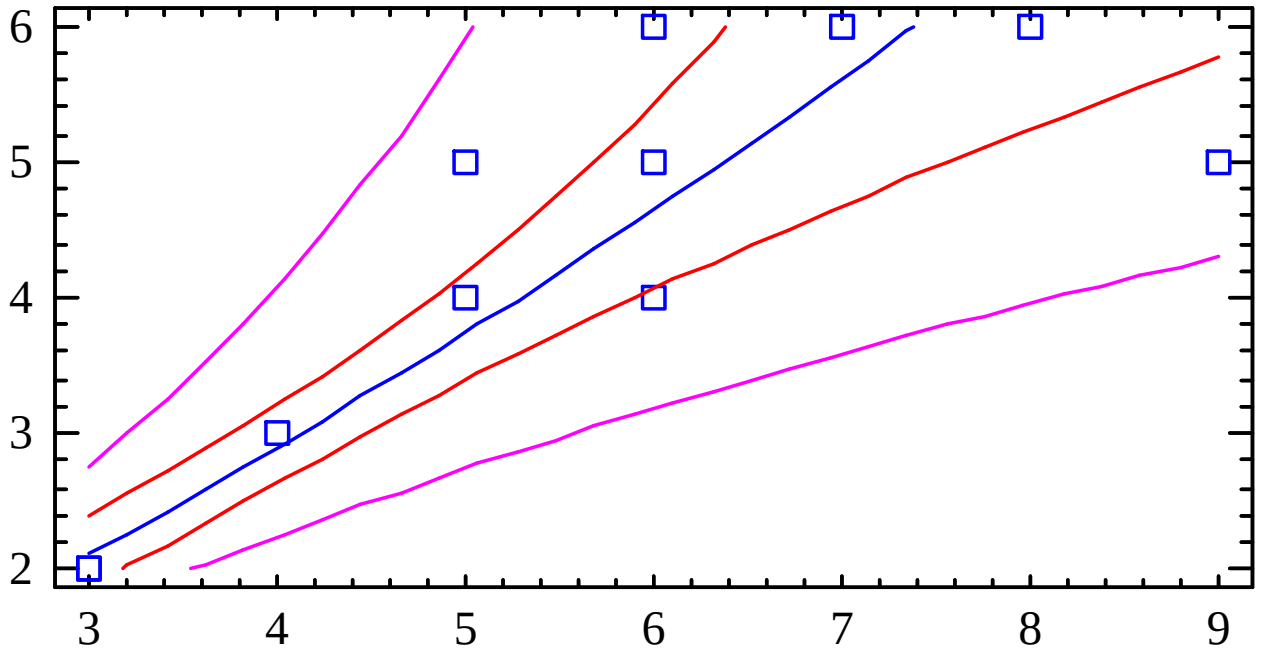
1) $g(x)=\ln(x) \rightarrow \tilde{Y} = a*\ln(X)+b$
2) $g(x)=x^2 \rightarrow \tilde{Y} = a+b*X^2$
3) $f(y)=y^{1/2} \rightarrow \tilde{Y}^{1/2}=a+bX \rightarrow \tilde{Y} = (a+bX)^2$

Przykład:

Dane jak z poprzedniego przykładu

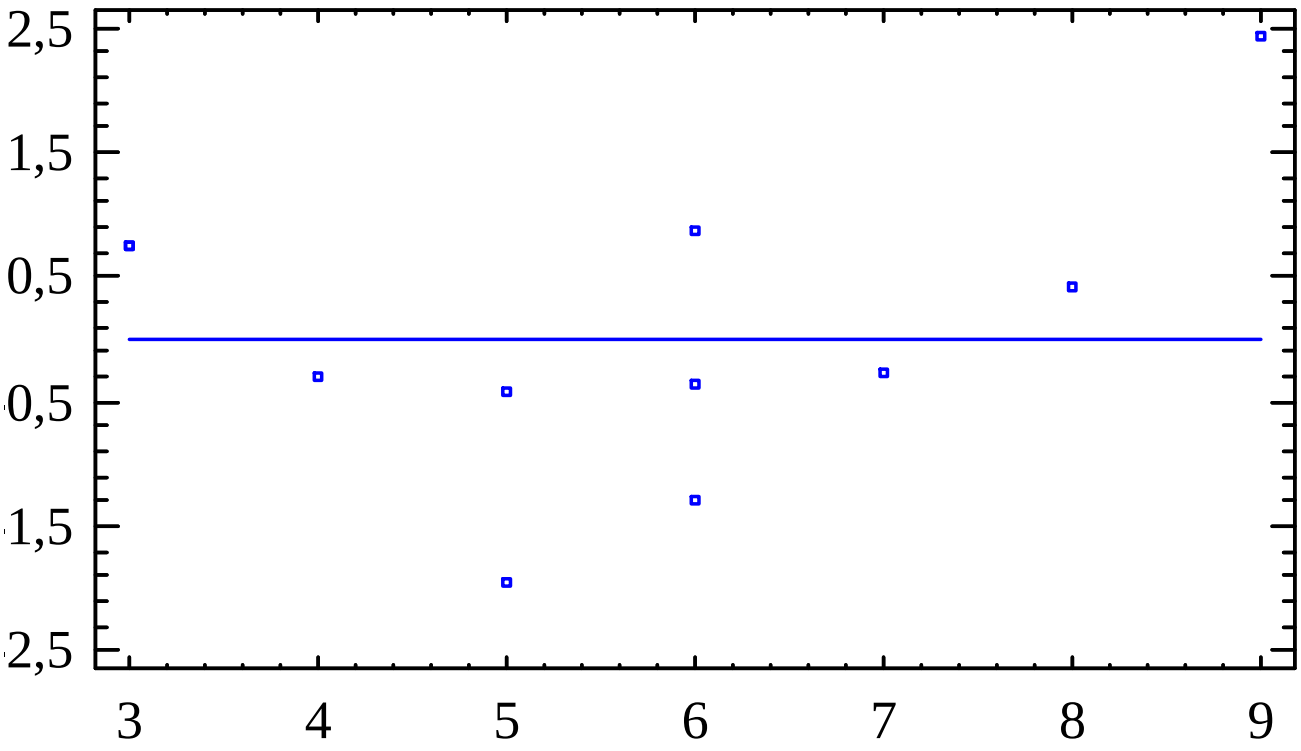
Przyjmujemy model: $Y = {}^{1}/(a + b*{}^{1}/_{X})$, tzn. $f(y)={}^{1}/_{y}$ i $g(x)={}^{1}/_{x}$.

Współczynnik determinacji dla tego modelu wynosi 90,1% przy a=1,56 i b=–0,04.

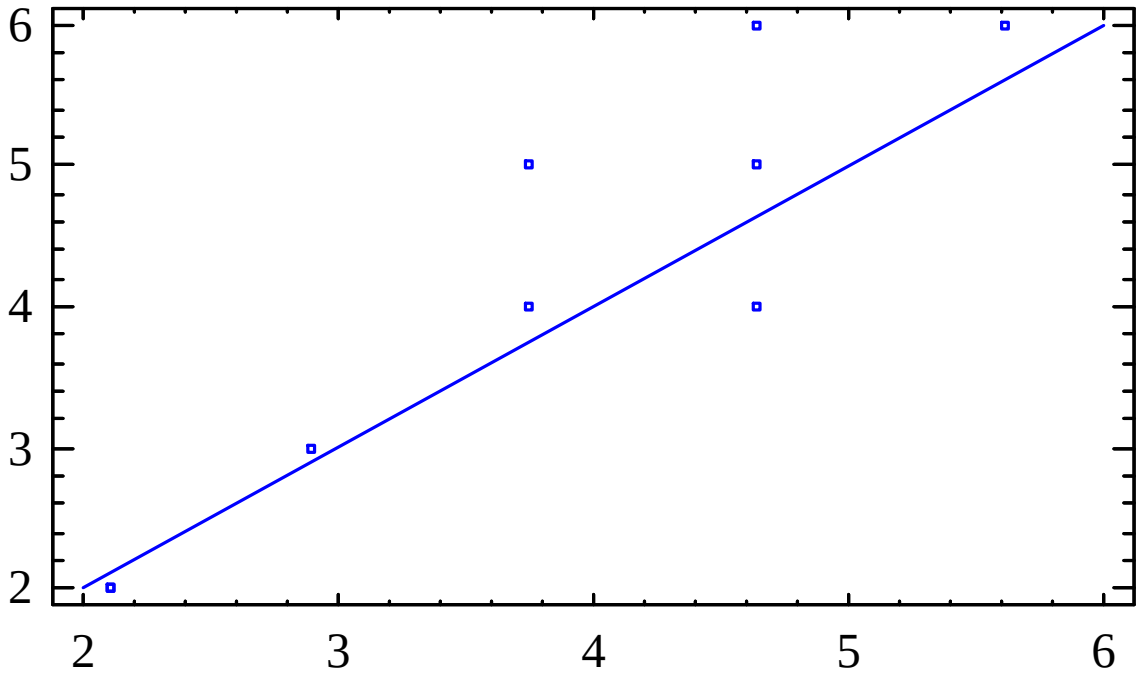


Na osi X znajduje się zmienna niezależna X, na osi Y zmienna zależna Y. Punkty obserwowane są rozmieszczone losowo w obrębie obszaru przewidywanego przez model, również błędy są rozłożone losowo

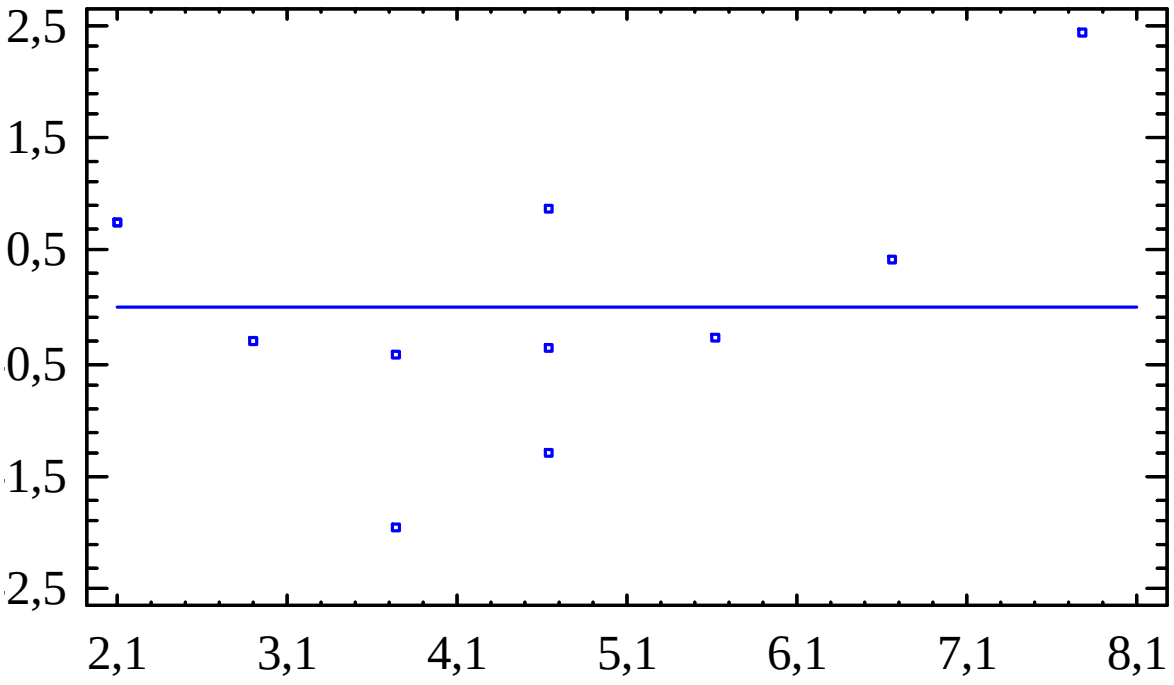
Wykres przedstawia odchylenia od krzywej regresji widocznej na poprzednim wykresie.



Wykres przedstawia przewidywaną przez model wartość (OX) i obserwowaną dla niej wartość.



Wykres przedstawia przewidywaną przez model wartość (OX) i odchylenia obserwowanych dla niej wartości.



Korelacja rangowa

Fakt zależności nieliniowej ale **monotonicznej** można wykryć za pomocą współczynnika korelacji rangowej (współczynnika Spearmana).

Rangowanie obiektów: Rangą nazywa się kolejność danego obiektu w próbie. „1” można przypisać obiektowi z największą wartością (porządek malejący) lub z najmniejszą wartością (porządek rosnący). Dla większości zastosowań przyjęcie porządku malejącego czy rosnącego nie ma znaczenia.

Powtarzające się wartości zmiennych — rangi wiązane. Jeżeli mamy próbę: 5,5; 6,3, 6,3,7,1 to jak należy przypisać rangi? 1; 2; 3; 4 (uwzględniając kolejność obserwacji)? 1; 2; 2; 4 (jak w sporcie)? czy **1; 2,5; 2,5; 4**?

Współczynnik ten jest odporny (*robust*) na obserwacje odstające i nienormalność rozkładu cech.