Regresja liniowa

W sytuacji, gdy obserwowana jest zmienna dwuwymiarowa (X,Y) stawiamy pytanie, czy występuje związek prostoliniowy pomiędzy tymi zmiennymi, czy jedna z nich jest zmienną zależną, a druga zmienną niezależną. Zatem na podstawie obserwacji (x_1,y_1), (x_2,y_2), …, (x_n,y_n) wyznaczana jest wartość współczynnika korelacji **Pearsona** $\mathbf{r_{xy}}$ (wzory na końcu materiałów).

- Stawiamy hipotezę: $\mathbf{H_0 : \rho = 0,}$
  $\mathbf{H_1 : \rho \neq 0}$

i na przyjętym poziomie istotności sprawdzamy, czy badane zmienne są skorelowane. Odrzucenie hipotezy zerowej implikuje wyznaczenie **prostej regresji**. Jeżeli dla każdej wartości X zmienna Y ma rozkład normalny z jednakową (nieznaną) wariancją to prostą regresji można zapisać w postaci

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

przedstawiającej związek między badanymi zmiennymi. Możemy też zweryfikować hipotezę dla współczynnika regresji, wyznaczyć krzywe ufności, obliczyć wartość miary dopasowania prostej regresji do punktów eksperymentalnych jaką jest współczynnik determinacji R^2 (wzory na końcu opracowania).

Zadanie 1. Badając zanieczyszczenie terenów wokół pewnego obiektu przemysłowego, odsłonięto siedem profili glebowych. W powierzchniowej warstwie badanych profili zawartości ołowiu i cynku (w mg/kg) przedstawiały się następująco:

ołów (X)	355	190	345	316	269	210	275
cynk (Y)	82	53	93	82	67	46	80

Przy prawdziwości założeń analizy regresji:

a) Oblicz i zinterpretuj współczynnik korelacji między cechami X i Y.

b) Sprawdź hipotezę o braku korelacji między zawartością ołowiu i cynku w powierzchniowej warstwie badanych profili. Przyjmij poziom istotności 0,05.

c) Wyznacz równanie regresji liniowej zawartości cynku względem zawartości ołowiu w powierzchniowej warstwie badanych profili. Zinterpretuj współczynnik regresji.

d) Oblicz i zinterpretuj współczynnik determinacji.

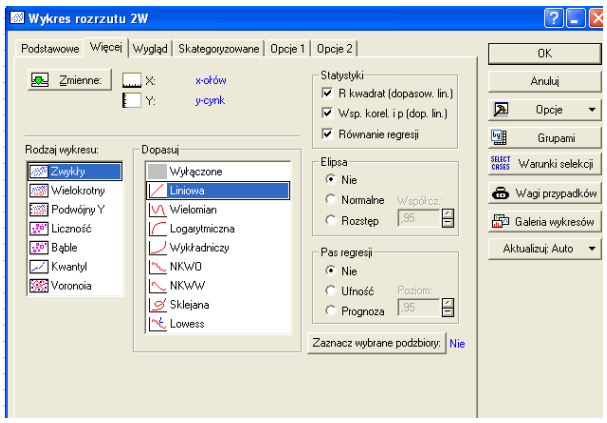
e) Na poziomie istotności $\alpha = 0{,}05$ zweryfikuj, za pomocą analizy wariancji, hipotezę o braku regresji liniowej zawartości cynku względem zawartości ołowiu.

f) Wyznacz przewidywaną zawartość cynku w powierzchniowej warstwie, gdy zawartość ołowiu wynosi 270 (mg/kg).

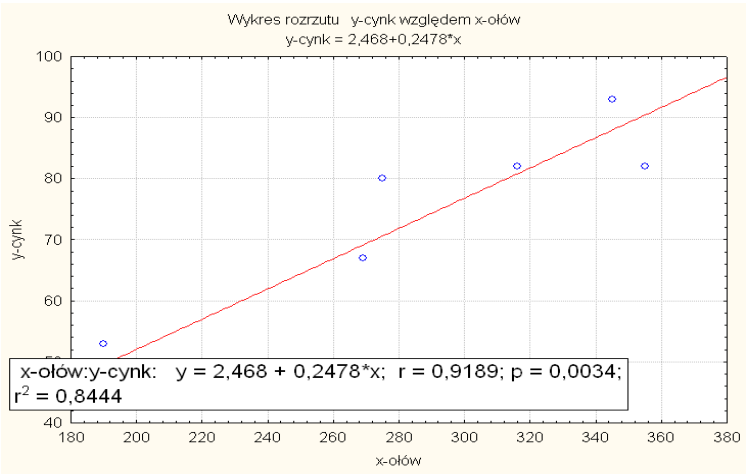
g) Zbuduj 95 % przedział ufności dla oczekiwanej zawartości cynku, jeśli zawartość ołowiu wynosi 270 mg/kg.

Za pomocą programu STATISTICA
Dane należy wpisać w dwóch kolumnach

	1 x-ołów	2 y-cynk
1	355	82
2	190	53
3	345	93
4	316	82
5	269	67
6	210	46
7	275	80



Zapoznamy się najpierw z wykresem zależności Y (zawartości cynku) od X (zawartości ołowiu). Z menu wybieramy **Wykresy** → **Wykresy rozrzutu** → w oknie **Zmienne** wybieramy odpowiednie zmienne → **OK** → **Więcej** → **Dopasuj** → wybieramy **Liniowa** oraz **R kwadrat, wsp. korel. i p, równanie regresji** → **OK**. Otrzymamy wykres wraz z wynikami.



Odpowiedzi:

a) Współczynnik korelacji r = 0,9189 jest dodatni, czyli korelacja między cechami jest dodatnia: im większa jest zawartość ołowiu w glebie, tym większa jest zawartość cynku.

b) $H_0 : \rho = 0$ korelacja między cechami jest nieistotna (nie ma korelacji)

$H_1 : \rho \neq 0$ korelacja między cechami jest istotna (jest korelacja)

Ponieważ p = 0,00344 < $\alpha = 0{,}05$, zatem hipotezę H_0 odrzucamy na poziomie istotności $\alpha = 0{,}05$ i stwierdzamy, że korelacja między zawartościami ołowiu i cynku w glebie jest istotna.

c) Prosta regresji wyraża się wzorem $y = 2{,}468 + 0{,}2478\,x$. Jeżeli zawartość ołowiu wzrośnie o jednostkę, czyli o 1 mg/kg, to przeciętna zawartość cynku wzrośnie o 0,25 mg/kg.

d) Współczynnik determinacji $R^2 = r^2 \cdot 100\%$, czyli u nas $R^2 = 0{,}8444 \cdot 100\% = 84{,}44\%$. Oznacza to, że zmienna X w ponad 84% ma wpływ na wartość zmiennej Y.

e) $H_0 : \beta_1 = 0$ regresja liniowa jest nieistotna

$H_1 : \beta_1 \neq 0$ regresja liniowa jest istotna

Aby zweryfikować tę hipotezę należy z menu **Statystyka** wybrać opcję **Regresja wieloraka**, wskazać zmienną zależną Y i zmienną niezależną X → OK → OK. W oknie **Wyniki regresji wielorakiej**, należy wybrać przycisk **Więcej**, a następnie opcję **ANOVA (sum. dobroć dopasow.).** Na ekranie pojawi się tabela analizy wariancji i wyniki testu F:

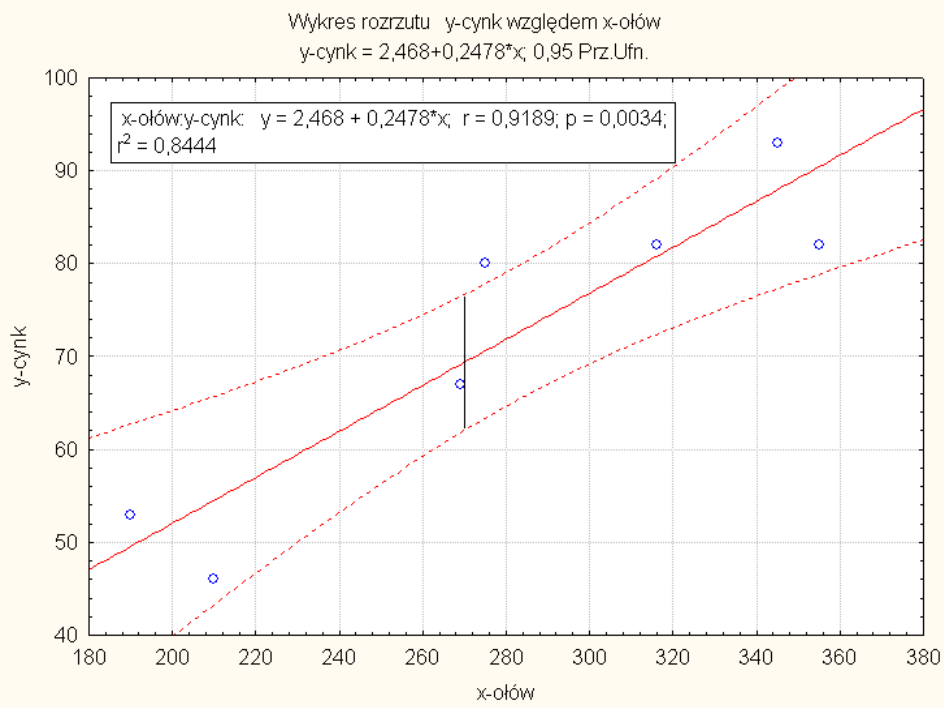
	Analiza wariancji ; DV: y-cynk (Arkusz1)				
Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	poziom p
Regres.	1491,866	1	1491,866	27,12567	0,003443
Reszta	274,992	5	54,998		
Razem	1766,857				

	Obliczanie wartości (Arkusz1) zmiennej: y-cynk		
Zmienna	Waga B	Wartość	Waga B *Wartość
x-ołów	0,247818	270,0000	66,91092
W. wolny			2,46804
Przewidyw.			69,37896
-95,0%GU			62,07050
+95,0%GU			76,68742

Ponieważ p = 0,00344 < $\alpha = 0{,}05$, to na poziomie istotności $\alpha = 0{,}05$ hipotezę zerową H_0 odrzucamy i stwierdzamy, że regresja liniowa zawartości cynku względem zawartości ołowiu jest istotna.

f) Wracamy do okna **Wyniki regresji wielorakiej**, wybieramy przycisk **Reszty**, **założenia, predykcja**, a następnie **Predykcja zmiennej zależnej** (przy aktywnym poleceniu **Oblicz granice ufności**).
W oknie **Określ wartości zmiennych niezależnych** wpisujemy liczbę 270 i naciskamy OK. Otrzymujemy ocenę punktową prognozy oraz poszukiwane granice ufności wartości prognozowanej.
Z tabeli wynika, że jeśli zawartość ołowiu wynosi 270 mg/kg, to przewidywana zawartość cynku wynosi 69,38 mg/kg.

g) Z tej samej tabeli odczytujemy lewy i prawy kraniec przedziału ufności.
Z prawdopodobieństwem 95 % stwierdzamy, że przedział (62,07; 76,69) pokrywa oczekiwaną zawartość cynku (w mg/kg) przy zawartości ołowiu 270 mg/kg.
Dodatkowo można wyznaczyć obszar ufności dla oczekiwanej zawartości cynku przy różnych zawartościach ołowiu i ew. zilustrować („ręcznie") przedział ufności dla x = 270. Postępujemy podobnie jak w punkcie a), dodatkowo zaznaczając **Pas regresji → ufność**.



Zad. 2. Na terenie byłego województwa konińskiego badano zmniejszenie się emisji pyłu (w t/rok) po zamontowaniu instalacji mokrego odpylania na kominach największych zakładów. Otrzymano dane:

liczba zamontowanych instalacji odpylania (X)	1	2	3	3	4	5
zmniejszenie emisji pyłu (Y)	5,8	6,1	8,4	9,2	9,3	10,4

Przy prawdziwości założeń analizy regresji:
a) Wyznacz równanie regresji liniowej zmniejszenia emisji pyłów względem liczby instalacji mokrego odpylania. Zinterpretuj współczynnik regresji.

Odp. $y = 4{,}48 + 1{,}24x$;

b) Na poziomie istotności 0,05 zweryfikuj hipotezę o braku regresji liniowej zmniejszenia emisji pyłów względem liczby instalacji mokrego odpylania.
c) Określ przewidywane zmniejszenie emisji pyłów, gdy liczba instalacji wyniesie 3.

Odp. 8,2

Zad. 3. Dział marketingu pewnej firmy analizował związek między wielkością sprzedaży swych produktów (w tys. sztuk) a liczbą współpracujących z zakładem hurtowni. Otrzymano dane:

liczba hurtowni (X)	1	2	3	4	5	6	7	8	9	10
wielkość sprzedaży (Y)	5,8	6,1	8,4	9,2	9,3	10,4	12,9	14,6	19,1	22,8

Przy prawdziwości założeń analizy regresji:
a) Oblicz i zinterpretuj współczynnik korelacji między wielkością sprzedaży produktów a liczbą współpracujących z zakładem hurtowni. Odp. r = 0,95
b) Wyznacz równanie regresji liniowej wielkości sprzedaży produktów względem liczby współpracujących z zakładem hurtowni. Zinterpretuj współczynnik regresji.

Odp. $y = 2{,}29 + 1{,}74x$

c) Zweryfikuj hipotezę o istotności regresji liniowej wielkości sprzedaży względem liczby współpracujących z zakładem hurtowni. Przyjmij poziom istotności $\alpha = 0.05$. Odp. p=000031
d) Określ przewidywaną wielkość sprzedaży, gdy liczba hurtowni wyniesie 6. Odp. 12,73

Zad. 4. Badano zawartość tlenu rozpuszczonego w wodzie destylowanej (cecha Y w mgO_2/dm^3) w zależności od temperatury (cecha X w ^oC). Uzyskano dane:

temperatura (X)	5	7	8	11	13	14	16	17	20	21
zawartość tlenu (Y)	12,9	13,6	11,9	11	11,2	11,9	10	11,7	8,8	8,9

Przy prawdziwości założeń analizy regresji:
a) Wyznacz równanie regresji liniowej zawartości tlenu rozpuszczonego w wodzie destylowanej względem temperatury. $y = 14{,}5 - 0{,}25x$
b) Jakiej zmiany zawartości tlenu w wodzie możemy się spodziewać, gdy temperatura wody wzrośnie o 1 ^oC?
c) Na poziomie istotności 0,05 zweryfikuj hipotezę o braku regresji liniowej zawartości tlenu rozpuszczonego w wodzie destylowanej względem temperatury. Odp: p=0,001145
d) Określ przewidywaną zawartość tlenu, gdy temperatura wynosi 12 ^oC. Odp. 11,5
e) Obliczyć i zinterpretować współczynnik determinacji. Odp. R^2 = 75,26%
f) Zbudować 95 % przedział ufności dla prognozowanej zawartości tlenu przy temperaturze 10 ^0C. Odp. (11,27955; 12,70714)

Zad. 5. Badano zależność między roczną wielkością wytworzonych odpadów w Polsce w mln ton wg GUS a ilością odpadów wykorzystanych wtórnie w ciągu roku w mln ton. Uzyskano następujące dane:

Dla X (wytworzone odpady):	120,8	122,7	124,6	124,4	133,2
Dla Y (wykorzystane odpady) :	65,6	66,9	69,5	80,1	91,7

Przy prawdziwości założeń analizy regresji:

a) Oblicz i zinterpretuj współczynnik korelacji między wielkością wykorzystanych odpadów a ilością wytworzonych odpadów.

Odp. r = 0,92;

b) Oszacuj prostą regresji wielkości wykorzystanych odpadów względem wytworzonych odpadów.

$y = -193{,}72 + 2{,}15x$

c) Określ przewidywaną wielkość wykorzystanych odpadów, gdy ilość wytworzonych odpadów wynosi 130 mln ton. Odp. 85,78

d) Zbuduj 95 % przedział ufności dla oczekiwanej wielkości wykorzystanych odpadów, gdy wielkość wytworzonych odpadów wynosi 125 mln ton.

Zad. 6. Producent napojów gazowanych dla sprawdzenia, czy istnieje związek między wielkością zamówień hurtowni a temperaturą dobową zgromadził dane dotyczące zamówień i temperatury dla wybranych 10 dni czerwca:

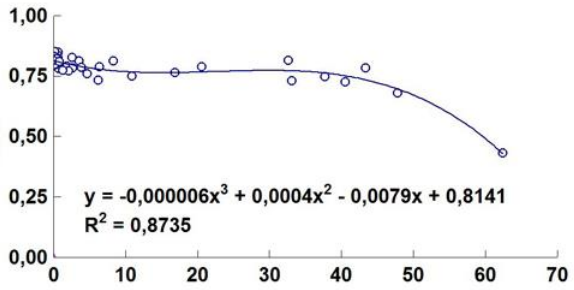
Temperatura dobowa w °C (X)	18	25	30	34	18	19	21	24	29	17
Zamówienia napojów gazowanych (w tys. sztuk) (Y)	7,2	12	12	15	9,4	4,5	10	12	13	7

Przy prawdziwości założeń analizy regresji:
a) Wyznacz równanie regresji liniowej zamówień hurtowni względem temperatury dobowej. Zinterpretować współczynnik regresji.
b) Na poziomie istotności $\alpha = 0{,}05$ zweryfikuj, za pomocą analizy wariancji, hipotezę o braku regresji liniowej wielkości zamówień napojów względem temperatury.

Regresja wielomianowa (metoda krokowa wsteczna)

Na *n* elementach próbki losowej pobranej z populacji normalnej albo w przybliżeniu normalnej obserwowana jest zmienna dwuwymiarowa (*X*,*Y*). Stawiamy pytanie, czy występuje związek wielomianowy pomiędzy tymi zmiennymi i którego stopnia.

Na podstawie obserwacji (x_1y_1), (x_2y_2), …, (x_ny_n) wyznaczany jest diagram korelacyjny, który ułatwi podjęcie decyzji dotyczącej stopnia wielomianu $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k$.



Przy spełnieniu założeń analizy regresji wielomianowej, wyznaczane jest równanie regresji wielomianowej obranego stopnia. W przypadku istnienia silnego związku między zmienną zależną *Y* a zmienną niezależną *X*, ważne jest by współczynnik regresji przy najwyższej potędze zmiennej *X* w wyznaczonym równaniu regresji wielomianowej był istotny. Nie zawsze jednak tak jest.

Dobór właściwego równania regresji wielomianowej można przeprowadzić stosując metodę krokową zstępującą (wsteczną).

Metoda krokowa zstępująca (wsteczna)
1. Wyznaczenie modelu wielomianowego o potencjalnie wysokim stopniu.
2. Wyznaczenie współczynnika determinacji.
3. Weryfikacja hipotez dla cząstkowych współczynników regresji, przede wszystkim współczynnika regresji przy najwyższej potędze zmiennej *X* ($H_0 : \beta_k = 0; H_1 : \beta_k \neq 0$).
4. Po stwierdzeniu nieistotności współczynnika regresji przy najwyższej potędze zmiennej *X*, wyznacza się model wielomianowy stopnia mniejszego o jeden.
5. Powrót do punktu 2.

Postępowanie tak długo jest kontynuowane do uzyskania modelu wielomianowego o istotnym współczynniku regresji przy najwyższej potędze zmiennej *X*.

Zadanie 7

W literaturze przedstawionych jest wiele metod pozwalających na oszacowanie wartości przepływu wód w rzekach w przekrojach niekontrolowanych. Ze względu na silną zależność przepływów w rzekach polskich od sezonu roku, przyjęto, że badane będą zależności regresyjne dla każdego miesiąca oddzielnie. W marcu uzyskano następujące wyniki :

Powierzchnia zlewni w km^2	50	120	200	250	320	340	360	410	470	560	590	610	670	700
Wartość przepływu w m^3/s	2	5	10	13	16	17	16	20	24	27	26	30	31	34

Wyznacz model regresji krzywoliniowej stopnia 2. Zastosuj metodę krokową wsteczną do wyznaczenia ostatecznej postaci funkcji przedstawiającej związek wartości przepływu i powierzchni zlewni. Wyznacz krzywe ufności. Przyjmij α = 0,05.

Rozwiązanie:

STATISTICA: Wpisujemy dane w dwóch kolumnach.

Następnie postępujemy według schematu:

Statystyka → **Zaawansowane modele liniowe i nieliniowe** → **Ogólne modele regresji** → **Kreator analizy** → **OK** → Następuje ustalenie **zmiennych**: zmienna zależna (tutaj *wartość przepływu*) i predykatora ciągłego (tutaj *powierzchnia zlewni*) → **OK**. Przechodzimy do zakładki **Dostosowany układ międzygrupowy** → klikamy na pozycję w okienku '**Ciągłe**'. Poniżej opcji '**Wielom. do st.**' ustalamy stopień wielomianu i klikamy przycisk **Wielom. do st.**. W okienku '**Efekty w układzie międzygrupowym**' pojawia się nazwa zmiennej niezależnej (predykator ciągły) w pierwszej i kolejnych potęgach do wybranego stopnia → **OK** → **Wszystkie efekty**.

KROK 1

Interesuje nas przede wszystkim wynik weryfikacji hipotezy: $H_0 : \beta_2 = 0;$ przeciwko $H_1 : \beta_2 \neq 0$.

W skoroszycie mamy trzy tabele z wynikami:

Tabela 1: Jednowymiarowe testy istotności.

Efekt	Jednowymiarowe testy istotności dla Wartość przepływu w m3/s (Cw_10_dane) Parametryzacja z sigma-ograniczeniami Dekompozycja efektywnych hipotez SS	Stopnie swobody	MS	F	p
Wyraz wolny	0,50821	1	0,50821	0,44345	0,519182
Powierzchnia zlewni w km2	85,51787	1	85,51787	74,62025	0,000003
Powierzchnia zlewni w km2^2	1,25572	1	1,25572	1,0957	0,317667
Błąd	12,60645	11	1,14604		

Z tabeli tej odczytujemy, że współczynnik dla drugiej potęgi jest nieistotny.

Tabela 2: Oceny parametrów:

Efekt	Oceny parametrów (Arkusz1) Parametryzacja z sigma-ograniczeniami Wartość przepływu w m3/s Param.	Wartość przepływu w m3/s Bł. std.	Wartość przepływu w m3/s t	Wartość przepływu w m3/s p	-95,00% Gr.ufn.	+95,00% Gr.ufn.	Wartość przepływu w m3/s Beta (ß)	Wartość przepływu w m3/s Bł.Std.ß	-95,00% Gr.ufn.	+95,00% Gr.ufn.
Wyraz wolny	-0,724996	1,088710	-0,66592	0,519182	-3,12123	1,671238				
Powierzchnia zlewni w km2	0,053842	0,006233	8,63830	0,000003	0,04012	0,067560	1,127327	0,130503	0,840091	1,414562
Powierzchnia zlewni w km2^2	-0,000008	0,000008	-1,04676	0,317667	-0,00003	0,000009	-0,136606	0,130503	-0,423841	0,150630

W tej tabeli zawarta jest również informacja, że współczynnik dla drugiej potęgi jest nieistotny, a także podane są wartości oszacowanych cząstkowych współczynników regresji dla zmiennej *X* występującej w kolejnych potęgach oraz oszacowanie wyrazu wolnego.

Tabela 3: Test SS dla pełnego modelu względem reszt.

	Test SS dla pełnego modelu względem SS dla reszt (Arkusz1)										
Zależna Zm.	Wielokr. R	Wielokr. R2	Skorygow R2	SS Model	df Model	MS Model	SS Reszta	df Reszta	MS Reszta	F	p
Wartość przepływu w m3/s	0,994950	0,989925	0,988093	1238,608	2	619,3039	12,60645	11	1,146041	540,3855	0,000000

Z tej tabeli odczytujemy stopień dopasowania modelu do danych i istotność związku regresyjnego wyrażonego równaniem regresji wielomianowej tutaj stopnia drugiego.

KROK 2

Ponieważ współczynnik przy drugiej potędze jest nieistotny (nie odrzucono $H_0 : \beta_2 = 0$), zmniejszamy stopień modelu. W tym celu otwieramy ponownie okienko GRM-wyniki, klikamy **Zmień→ Dostosowany układ międzygr.** → z '**Efekty w układzie międzygrupowym**' usuwamy zmienną w drugiej potędze → **Ok** → **Wszystkie efekty**. W skoroszycie tworzą się ponownie trzy tabele z wynikami jak w kroku 1.

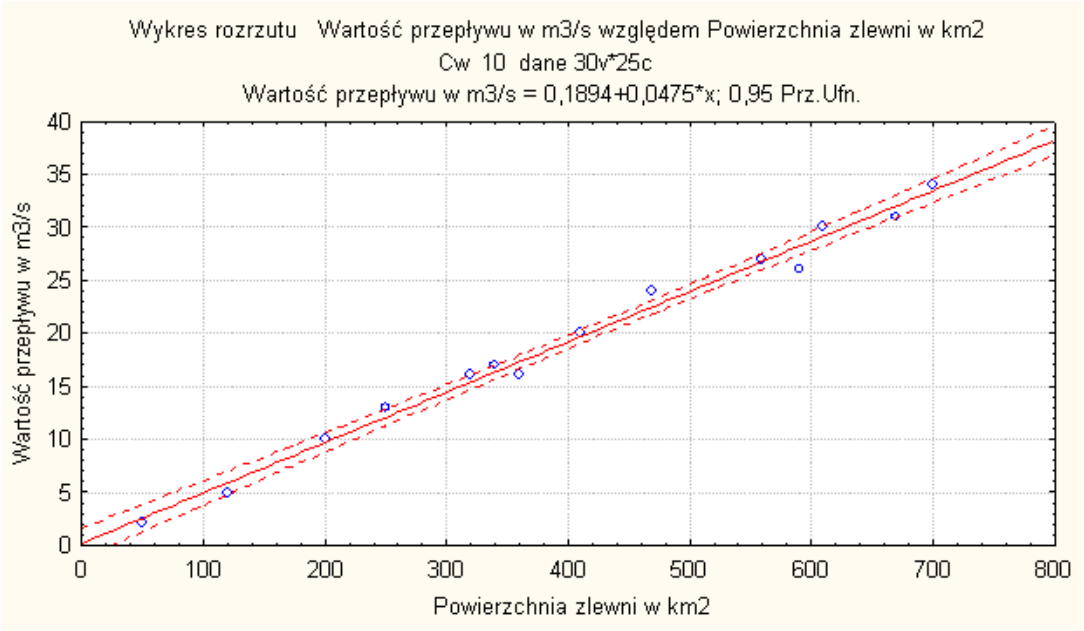
Interesuje nas wynik weryfikacji hipotezy $H_0 : \beta_1 = 0;$ przeciwko $H_1 : \beta_1 \neq 0$

	Oceny parametrów (Arkusz1) Parametryzacja z sigma-ograniczeniami									
Efekt	Wartość przepływu w m3/s Param.	Wartość przepływu w m3/s Bł. std.	Wartość przepływu w m3/s t	Wartość przepływu w m3/s p	-95,00% Gr.ufn.	+95,00% Gr.ufn.	Wartość przepływu w m3/s Beta (ß)	Wartość przepływu w m3/s Bł.Std.ß	-95,00% Gr.ufn.	+95,00% Gr.ufn.
Wyraz wolny	0,189431	0,652315	0,29040	0,776469	-1,23184	1,610705				
Powierzchnia zlewni w km2	0,047495	0,001451	32,72816	0,000000	0,04433	0,050657	0,994445	0,030385	0,928242	1,060648

	Test SS dla pełnego modelu względem SS dla reszt (Arkusz1)										
Zależna Zm.	Wielokr. R	Wielokr. R2	Skorygow R2	SS Model	df Model	MS Model	SS Reszta	df Reszta	MS Reszta	F	p
Wartość przepływu w m3/s	0,994445	0,988921	0,987998	1237,352	1	1237,352	13,86217	12	1,155181	1071,132	0,000000

Otrzymujemy ostatecznie równanie regresji liniowej pozwalające na oszacowanie wartości zmiennej y dla danego x, postaci: $y = 0{,}189 + 0{,}047x$

Krzywe ufności wyznaczamy wybierając z menu **Wykresy** → **Wykresy rozrzutu** → w oknie **Zmienne** wybieramy odpowiednie zmienne → **OK** → **Więcej** → **Dopasuj** → wybieramy **Liniowa** oraz **Pas regresji** zaznaczając **ufność** → **OK**



Zadanie 8.

Testowano możliwość przewidywania stężeń tlenu rozpuszczonego na odcinku Raby 445 na podstawie znanych wartości stężeń tlenu w dopływach: Niżowskim Potoku i Krzyworzece. Zastosowano metodę sieci neuronowych do wyznaczenia tego związku. Wyniki otrzymanych tą

metodą przewidywanych wartości stężeń obok zaobserwowanych wartości stężeń tlenu na odcinku Raby 445 przedstawiono w tabeli :

Stężenie tlenu w Rabie 445	1,1	2,3	3,5	4,1	5,3	6,8	7,2	8,3	9,3	10,6	11,7	12,4	13,2	14,4
Wartość oszacowania stężenia	1,2	2,2	3,2	4,3	5,3	6,5	7,3	8,4	9,2	10,2	11,4	12,3	13,3	14,4

Sprawdź zgodność oszacowań z rzeczywistymi zaobserwowanymi wartościami stężeń tlenu w Rabie poprzez wyznaczenie związku między oszacowaniem a danymi rzeczywistymi. Zacznij od stopnia 3. Jeżeli ostatecznie po zastosowaniu metody krokowej wstecznej uzyskasz prostą regresji będącą dwusieczną I ćwiartki układu współrzędnych to potwierdzisz zgodność oszacowań z wynikami rzeczywistymi. Wyznacz krzywe ufności. Przyjmij $\alpha = 0{,}001$.

Odp. $y=-0{,}0187+0{,}9933x \approx y=x$

	Oceny parametrów (Arkusz22) Parametryzacja z sigma-ograniczeniami									
Efekt	y Param.	y Bł. std.	y t	y p	-95,00% Gr.ufn.	+95,00% Gr.ufn.	y Beta (ß)	y Bł.Std.ß	-95,00% Gr.ufn.	+95,00% Gr.ufn.
Wyraz wolny	-0,018729	0,113233	-0,16540	0,871380	-0,265442	0,227983				
x	0,993305	0,012764	77,81852	0,000000	0,965494	1,021116	0,999011	0,012838	0,971040	1,026982

Zadanie 9.

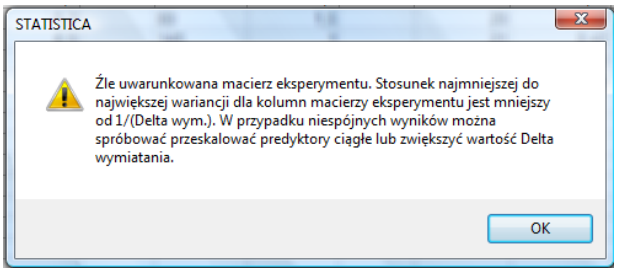
Dokonano pomiarów wielkości drgań pionowych gruntu powstałych w wyniku trzęsienia ziemi w różnej odległości od ogniska trzęsienia. Otrzymano wyniki (X – odległość od ogniska trzęsienia ziemi w km, Y – wielkość drgań pionowych gruntu w cm):

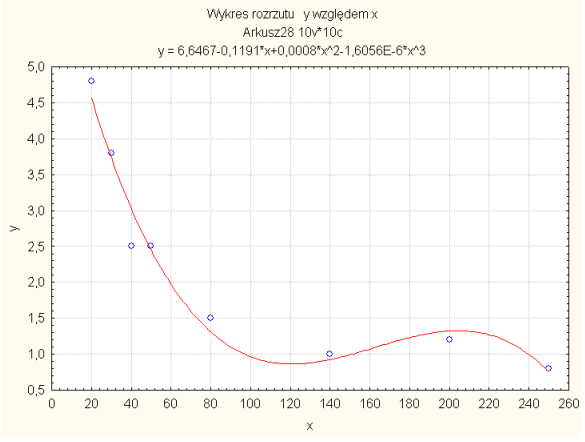
x	20	30	40	50	80	140	200	250
y	4,8	3,8	2,5	2,5	1,5	1,0	1,2	0,8

Wyznacz model regresji krzywoliniowej. Zastosuj metodę krokową wsteczną do wyznaczenia ostatecznej postaci funkcji przedstawiającej zależność wielkość drgań pionowych gruntu od odległości od ogniska trzęsienia ziemi. Rozpocznij od stopnia 3. Wykonaj wykres rozrzutu z dopasowaną krzywą. Wyznacz krzywe ufności. Przyjmij $\alpha = 0{,}05$.

Uwagi :

Zauważ, że dla stopnia 3 pojawia się komunikat.



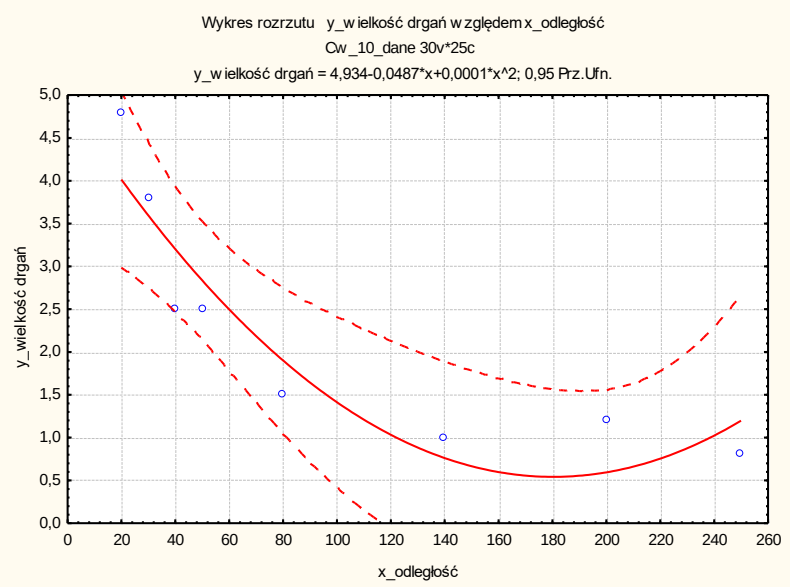


Oznacza to, że ustalono zbyt wysoki stopień wielomianu. Zignoruj ten komunikat przy stopniu 3 i wykonaj wykres rozrzutu y względem x. Uzyskasz wykres →

Z wykresu widać, że jeden punkt jest przyczyną zmiany wypukłości funkcji na wklęsłość. W badanym zagadnieniu szukania związku wielkości drgań pionowych gruntu powstałych w

wyniku trzęsienia ziemi od odległości od ogniska trzęsienia, zgodnie z pojawiającym się komunikatem stopień 2 jest właściwy do rozpoczęcia naszej analizy.

Odp.: $y = 4{,}934 - 0{,}0487x + 0{,}0001x^2$



Porównaj dopasowanie tej funkcji wielomianowej stopnia 2 z dopasowaniem do tych samych danych modelu hiperbolicznego (ćwiczenia : regresja liniowa i linearyzowana).

Zadanie 10

W pewnym doświadczeniu chemicznym obserwowano szybkość rozpuszczania się powłoki srebrnej w różnych temperaturach roztworu. Otrzymano wyniki (X – temperatura w stopniach, Y – szybkość rozpuszczania się powłoki w μ/sek):

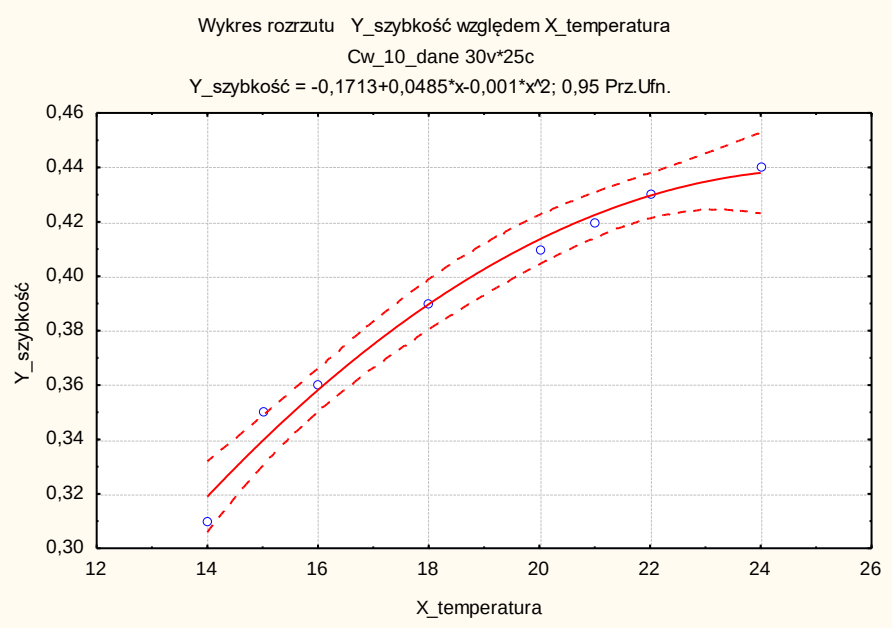
X	14	15	16	18	20	21	22	24
Y	0,31	0,35	0,36	0,39	0,41	0,42	0,43	0,44

Wyznacz model regresji wielomianowej. Zastosuj metodę krokową do wyznaczenia ostatecznej postaci funkcji przedstawiającej wpływ temperatury na szybkość rozpuszczania się powłoki. Rozpocznij od stopnia trzeciego. Wykonaj wykres rozrzutu z dopasowaną krzywą. Wyznacz krzywe ufności. Przyjmij $\alpha = 0{,}05$. Zinterpretuj współczynnik determinacji.

Odp. Krok 1: $H_0 : \beta_3 = 0$; $H_1 : \beta_3 \neq 0$, p = 0,199761 > $\alpha = 0{,}05$,

Krok 2: $H_0 : \beta_2 = 0$; $H_1 : \beta_2 \neq 0$ p = 0,013049 < $\alpha = 0{,}05$,

$y = -0{,}171327 + 0{,}048507\, x - 0{,}000963\, x^2$, R^2 = 98,4882%



Zadanie 11

W doświadczeniu badano dynamikę wzrostu traw ozdobnych pewnego wieloletniego gatunku. W szczególności mierzono średnicę kępy trzech wybranych roślin począwszy od połowy maja. Obserwacje prowadzono co 14 dni. Średnie średnice kęp S (w cm) w kolejnych okresach dwutygodniowych przedstawiały się następująco:

t	1	2	3	4	5	6	7
S	10	19	29	34	36	37,5	37,5

Wyznacz równanie regresji wielomianowej. Zastosuj metodę krokową do wyznaczenia ostatecznej dobrze dopasowanej do danych postaci modelu wielomianowego. Rozpocznij od stopnia 4. Wykonaj wykres rozrzutu z dopasowaną krzywą. Przyjmij $\alpha = 0{,}05$.

Rozwiązanie:

Krok 1: $H_0 : \beta_4 = 0;\ \ H_1 : \beta_4 \neq 0$,

Ponieważ $p_4 = 0{,}284816 > \alpha = 0{,}05$, to nie odrzucamy hipotezy zerowej i stwierdzamy, że współczynnik dla czwartej potęgi jest nieistotny i obniżamy stopień równania

Krok 2: $H_0 : \beta_3 = 0;\ \ H_1 : \beta_3 \neq 0$,

Ponieważ $p_3 = 0{,}490009 > \alpha = 0{,}05$, to nie odrzucamy hipotezy zerowej i stwierdzamy, że współczynnik dla trzeciej potęgi jest nieistotny i obniżamy stopień równania

Krok 3: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$

Ponieważ $p_2 = 0{,}000499 < \alpha = 0{,}05$, to odrzucamy hipotezę zerową i stwierdzamy, że współczynnik dla drugiej potęgi jest istotny.

Odp. Zależność średnicy (w cm) kępy traw ozdobnych od czasu (w tygodniach) jest postaci $S = -2{,}429 + 13{,}423t - 1{,}113t^2$. Współczynnik determinacji R^2 = 99,42% zatem otrzymany model jest w 99,42% dopasowany do danych.

Zadanie 12.

W badaniach nad stopniem skażenia gleby wokół pewnej huty pobrano próbki gleby z warstwy wierzchniej i czterech poziomów genetycznych w dwóch odkrywkach Uzyskano następujące oznaczenia cynku:

Głębokość (w cm)	2	10	41	70	90	2	15	52	100	110
Zn (w mg/kg)	172	88	72	72	60	136	136	80	64	45

Wyznacz równanie regresji wielomianowej. Zastosuj metodę krokową do wyznaczenia ostatecznej dobrze dopasowanej do danych postaci modelu wielomianowego. Rozpocznij od stopnia drugiego. Wykonaj wykres rozrzutu z dopasowaną krzywą. Przyjmij $\alpha = 0{,}05$.

Odp.: $y = 133{,}6964 - 0{,}83735\,x$

Zadanie 13.

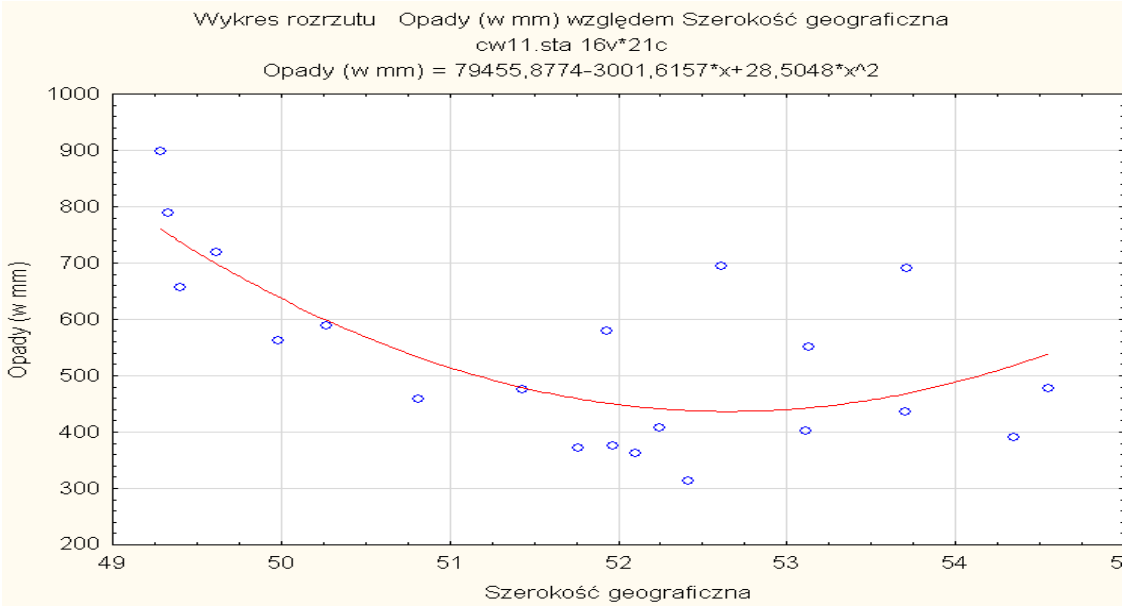
Suma opadów w okresie wegetatywnym od marca do października w niektórych miejscowościach (średnie z dwudziestu lat) oraz szerokość geograficzna, długość geograficzna i wzniesienie nad poziomem morza tych miejscowości są następujące (plik cw4.sta).
Wyznacz równania regresji wielomianowej wyrażające sumę opadów oddzielnie kolejno jako funkcję szerokości, jako funkcję długości oraz jako funkcję wysokości nad poziomem morza. Zastosuj metodę krokową wsteczną do wyznaczenia ostatecznej dobrze dopasowanej do danych postaci równania regresji wielomianowej. Rozpocznij od stopnia drugiego. Wykonaj dla każdej pary zmiennych wykres rozrzutu z dopasowaną krzywą. Przyjmij $\alpha = 0{,}05$.

Uwaga : Należałoby dla rozpatrywanych danych wyznaczyć również równanie regresji wyrażające jak suma opadów jest determinowana jednocześnie poprzez szerokość i długość geograficzną oraz wysokość nad poziomem morza (będzie to tematem kolejnych zajęć).

Odp.
Szerokość geograficzna:

Krok 1: $H_0 : \beta_2 = 0;\quad H_1 : \beta_2 \neq 0$, p = 0,011049 < $\alpha = 0{,}05$,

$y = 79455{,}88 - 3001{,}62x + 28{,}5x^2$, R^2 = 52,27%

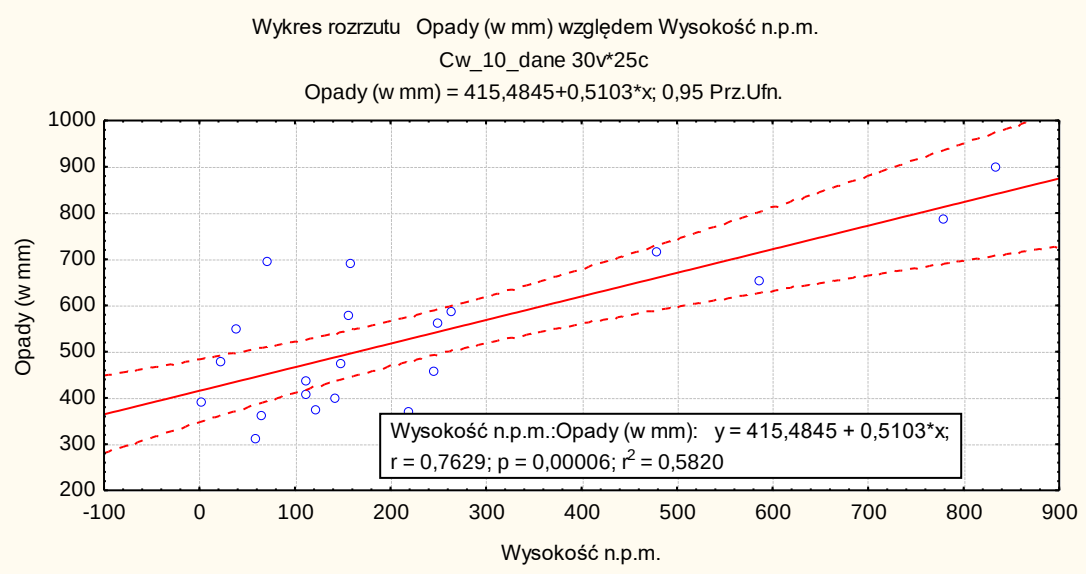


Długość geograficzna Zależność regresyjna nieistotna.
Wysokość n.p.m.

Krok 1: $H_0 : \beta_2 = 0;\quad H_1 : \beta_2 \neq 0$, p = 0,609179 > $\alpha = 0{,}05$,

Krok 2: $H_0 : \beta_1 = 0;\quad H_1 : \beta_1 \neq 0$ p = 0,000058 < $\alpha = 0{,}05$,

$y = 415{,}48 + 0{,}51x$, R^2 = 58,2%



Zadanie 14.

W pewnym nadleśnictwie, badając kondycję sosny dokonano wielu pomiarów i otrzymane obserwacje zapisano w pliku sosna.sta. Wyznacz równania regresji wielomianowej wyrażające objętość bielu oddzielnie kolejno jako funkcję wieku drzewa ($X1$), jako funkcję pierśnicy ($X2$), jako funkcję wysokości ($X3$), jako funkcję długości korony ($X4$), jako funkcję średnicy podstawy korony ($X5$), jako funkcję objętości korony ($X6$). Zastosuj metodę krokową wsteczną do wyznaczenia ostatecznej dobrze dopasowanej do danych postaci równania regresji wielomianowej. Rozpocznij od stopnia trzeciego. Wykonaj dla każdej pary zmiennych wykres rozrzutu z dopasowaną krzywą. Przyjmij $\alpha = 0{,}05$. Podaj współczynniki determinacji dla każdego przyjętego ostatecznie modelu.

Odp.

Objętość bielu (Y) jako funkcja wieku drzewa (X1)

Krok 1: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,532921 > $\alpha = 0{,}05$

Krok 2: $H_0 : \beta_1 = 0;\ \ H_1 : \beta_1 \neq 0$ p = 0,0000 < $\alpha = 0{,}05$

$y = -0{,}126972 + 0{,}007276\, x_1$, R^2 = 50,95%

Objętość bielu (Y) jako funkcja pierśnicy drzewa (X2)

Krok 1: $H_0 : \beta_3 = 0;\ \ H_1 : \beta_3 \neq 0$, p = 0,942128 > $\alpha = 0{,}05$

Krok 2: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,0000 < $\alpha = 0{,}05$

$y = 0{,}035883 - 0{,}005578\, x_2 + 0{,}000721\, x_2^2$; R^2 =83,32%

Objętość bielu (Y) jako funkcja wysokości drzewa (X3)

Krok 1: $H_0 : \beta_3 = 0;\ \ H_1 : \beta_3 \neq 0$, p = 0,352249 > $\alpha = 0{,}05$

Krok 2: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,0000 < $\alpha = 0{,}05$

$y = 0{,}515797 - 0{,}070855\, x_3 + 0{,}002901\, x_3^2$; R^2 =80,43%

Objętość bielu (Y) jako funkcja długości korony (X4)

Krok 1: $H_0 : \beta_3 = 0;\ \ H_1 : \beta_3 \neq 0$, p =0,571127> $\alpha = 0{,}05$

Krok 2: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,021402 < $\alpha = 0{,}05$

$y = -0{,}009896 + 0{,}017567\, x_4 + 0{,}004816\, x_4^2$; R^2 =65,60%

Objętość bielu (Y) jako funkcja średnicy podstawy korony (X5)

Krok 1: $H_0 : \beta_3 = 0;\ \ H_1 : \beta_3 \neq 0$, p = 0,449826 > α = 0,05

Krok 2: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,466011 > α = 0,05

Krok 3: $H_0 : \beta_1 = 0;\ \ H_1 : \beta_1 \neq 0$ p = 0,0000 < α = 0,05

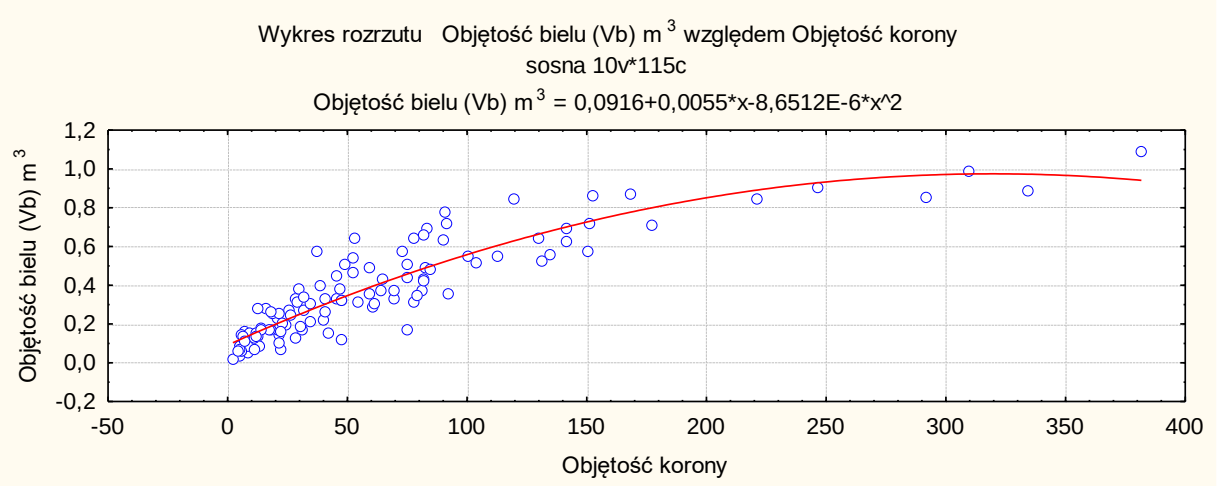
$y = -0191329 + 0{,}135082\, x_5$; R^2 =79,24%

Objętość bielu (Y) jako funkcja objętości korony (X6)

Krok 1: $H_0 : \beta_2 = 0;\ \ H_1 : \beta_2 \neq 0$, p = 0,0000 < α = 0,05

$y = 0{,}091624 + 0{,}005528\, x_6 - 0{,}000009\, x_6^2$; R^2 =82,41%

Jeden z sześciu wykresów:



ANALIZA KORELACJI I REGRESJI LINIOWEJ

Obserwacje (x_i,y_i) zmiennej losowej dwuwymiarowej (X,Y)

WSPÓŁCZYNNIK KORELACJI

$$r_{xy} = \frac{\hat{s}_{xy}}{\hat{s}_x \cdot \hat{s}_y}$$

przy czym

$-1 \le r_{xy} \le 1$

$$\hat{s}_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n}\right)$$

$$\hat{s}_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right), \qquad \hat{s}_x = \sqrt{\hat{s}_x^2}$$

$$\hat{s}_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right), \qquad \hat{s}_y = \sqrt{\hat{s}_y^2}$$

Weryfikacja hipotezy, że zmienne X i Y nie są skorelowane

Stawiamy hipotezę:	Statystyka testowa	Obszar krytyczny :
$H_0 : \rho = 0, \quad H_1 : \rho \neq 0$	$t = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \cdot \sqrt{n-2}$	$\lvert t\rvert > t_{\alpha,n-2}$

REGRESJA LINIOWA

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

dla $<x_{min}, x_{max}>$

$$\hat{\beta}_1 = \frac{\hat{s}_{xy}}{\hat{s}_x^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Weryfikacja hipotezy, że dla zmiennych X i Y
związek wyrażony za pomocą prostej regresji nie jest istotny

Stawiamy hipotezę:	Statystyka testowa	Obszar krytyczny :
$H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$	$t = \frac{\hat{\beta}_1}{\hat{s}_y \cdot \sqrt{1-r_{xy}^2}} \cdot \hat{s}_x \cdot \sqrt{n-2}$	$\lvert t\rvert > t_{\alpha,n-2}$

Miara dopasowania – współczynnik determinacji	$R^2 = r_{xy}^2 \cdot 100\%$	

Krzywe ufności

$$g_1(x_i) = y(x_i) - t_{\alpha,n-2} \cdot \sqrt{\frac{\hat{s}_y^2(1-r_{xy}^2)}{\hat{s}_x^2(n-2)}\left[s_x^2 + (x_i - \bar{x})^2\right]}$$

$$g_2(x_i) = y(x_i) + t_{\alpha,n-2} \cdot \sqrt{\frac{\hat{s}_y^2(1-r_{xy}^2)}{\hat{s}_x^2(n-2)}\left[s_x^2 + (x_i - \bar{x})^2\right]}$$