Ćwiczenie: Wybrane zagadnienia z korelacji i regresji.

W statystyce stopień zależności między cechami można wyrazić wg następującej skali:

Skala Guillforda

Przedział	Zależność	Współczynnik
[0,00±0,20)	Słaba	Prawie nic nieznaczący
[±0,20±040)	Niska	Wyraźna, ale słaba
[±0,40±0,70)	Umiarkowana	Rzeczywisty
[±0,70±0,90)	Wysoka	Znaczny
[±0,90±1,00]	Bardzo wysoka	Pewny

Istotność korelacji – weryfikacja hipotezy o niezależności cech. Polega ona na obliczeniu t_0 i porównaniu go z t_{tab}. Statystykę t_0 obliczamy w przypadku prób mniejszych od 122. Mając do czynienia z próbami liczącymi 122 i więcej stosujemy test z.

$$t_0 = r_{xy} * \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \qquad z_0 = r_{xy} * \frac{\sqrt{n}}{\sqrt{1-r_{xy}^2}}$$

t_{tab} odczytujemy z tabeli testu t dla poziomu istotności 0,05 i 0,01 oraz dla liczby stopni swobody równej n – 2.

Jeżeli $t_0 \geq t_{tab}$ to korelacja jest istotna statystycznie. Jeżeli $t_0 < t_{tab}$ to korelacja jest nieistotna statystycznie. Istotność korelacji jest liczona po to, aby sprawdzić czy zależność jaką stwierdzono w próbie będzie miała miejsce również w populacji, z której próba ta pochodzi.

Hipotez zerowa: $\mathbf{H_0}$: $\rho = 0$, zaś alternatywna $\mathbf{H_1}$: $\rho \neq 0$ ρ *(ro)*

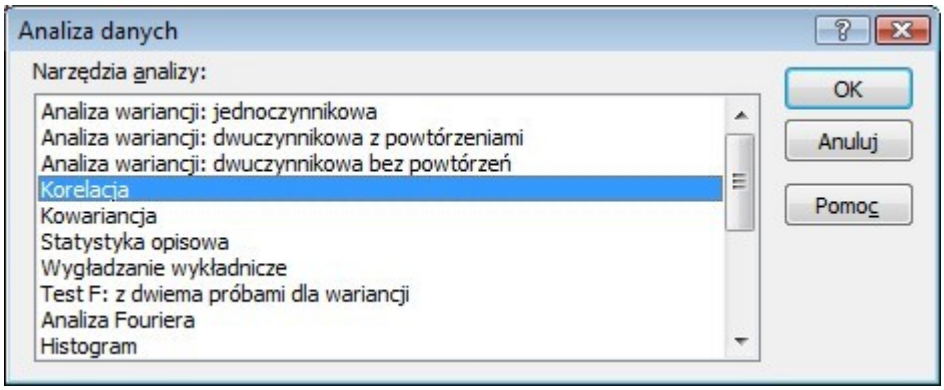
Funkcje pozwalające obliczyć współczynnik korelacji i regresji:
=wsp.korelacji(x2:x100;y2:y100) – współczynnik korelacji
=nachylenie(y2:y100;x2:x100) –współczynnik regresji
{=NACHYLENIE(znane_y ; znane_x)
Znane_y jest to tablica lub zakres komórek liczbowych zależnych punktów danych.
Znane_x jest to zbiór niezależnych punktów danych.}

I. Zadanie MS EXCEL:

1. Wykonaj wykres rozrzutu dla zmiennych: wzrost oraz długość stopy (stopy2011L.xls). Jaki charakter ma zależność między tymi zmiennymi (dodatni, ujemny)? Zawarte w **Zadaniu I** obliczenia wykonaj oddzielnie dla każdej płci.
2. Oblicz współczynnik korelacji pomiędzy wzrostem oraz długością stopy. Sprawdź czy jest to współczynnik istotny.
3. Oblicz współczynniki regresji między wzrostem i stopą. Załóż, że długość stopy jest zmienną zależną!
4. Do sporządzonego wykresu rozrzutu dołącz równanie regresji liniowej. Czy jest ono dobrze dopasowane do punktów w układzie współrzędnych?
   R-Square, R^2 (współczynnik determinacji) – informacja o tym, w jakim stopniu równanie regresji wyjaśnia zmienność zmiennej zależnej. To jest inaczej kwadrat współczynnika korelacji.
   R^2 przyjmuje wartości od 0 do 1 (0-100%). $R^2 = \frac{\sum y_p^2}{\sum y^2}$

II. Przykład MS EXCEL

1. Oblicz współczynniki korelacji Pearsona między mięsnością tuczników oraz grubością słoniny i schabu w dwóch punktach pomiaru (**tuczniki_ZAD_KOR.xls**). W tym celu zastosuj narzędzie Analiza danych => Korelacja.



W okienku Korelacja podajemy zakres zmiennych, które chcemy ze sobą korelować. Rozsądnym posunięciem będzie zaznaczenie również nazw kolumn. Należy jednak przy tym pamiętać, że konieczne jest w tej sytuacji zaznaczenie pola wyboru Tytuły w pierwszym wierszu. Decydujemy się na umieszczenie wyników w nowym arkuszu – klikamy na pole opcji **Nowy arkusz**.

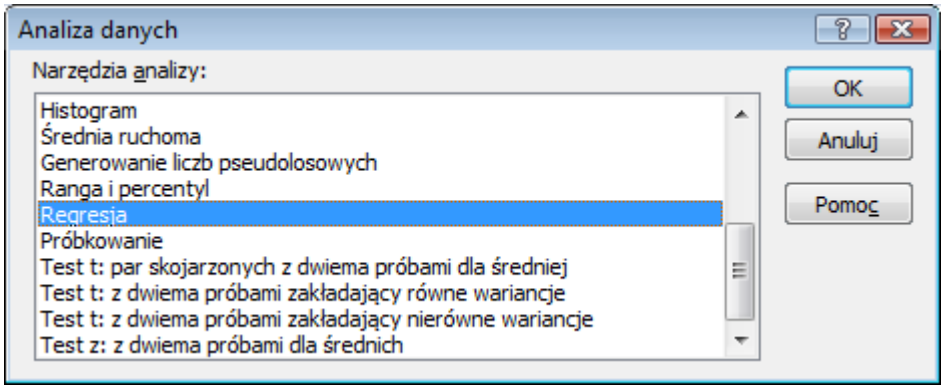
Efektem uruchomionej analizy jest obliczenie współczynników korelacji między wszystkimi cechami. Ze względu na fakt, iż z pewnych względów interesują nas współczynniki korelacji między mięsnością a pomiarami słoniny i schabu te właśnie poddamy interpretacji.

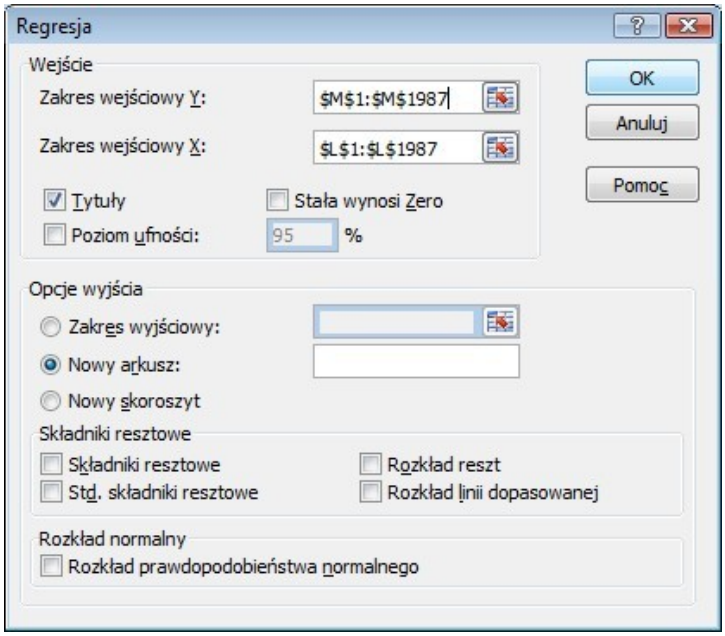
	A	B	C	D	E	F
1		*Miesnosc*	*Slonina1*	*Schab1*	*Slonina2*	*Schab2*
2	Miesnosc	1				
3	Slonina1	-0.94044	1			
4	Schab1	0.725753	-0.57027	1		
5	Slonina2	-0.90937	0.854517	-0.46616	1	
6	Schab2	0.702109	-0.49984	0.481008	-0.62345	1

Uzyskane wyniki pozwalają wnioskować, że wraz ze zwiększaniem się przekroju schabu wzrastała mięsność tuczników – obydwa obliczone współczynniki korelacji są dodatnie. Świadczą one jednocześnie, że zależności między tym cechami są wysokie, zaś same współczynniki korelacji – znaczne.

Negatywną (ujemną) zależność wykazano w przypadku mięsności tuczników oraz grubości słoniny w dwóch punktach. Posiłkując się skalą Guillforda można stwierdzić, że zależności są bardzo wysokie, a współczynniki korelacji pewne.

Jako kontynuację analizy zależności sporządzamy równanie regresji. Przyjmijmy, że Mięsność stanowi zmienną zależną, zaś Slonina1 zmienną niezależną. Wykorzystamy w tym celu narzędzie Analiza danych => Regresja.





Wyniki analiz. W załączonym ekranie pogrubiono sekcje poddane interpretacji. Przeprowadzona analiza regresji pozwoliła skonstruować równanie liniowe postaci: **Mięsność = 67,988 – 0,831Slonina1**. Obliczony współczynnik determinacji świadczy o tym, że model matematyczny dobrze opisuje zmienność mięsności. Wykonana jednocześnie analiza wariancji potwierdza, że skonstruowane równanie regresji jest istotne. Kolejne postępowanie statystyczne, tj. test t-Studenta dowodzi, że zarówno wyraz wolny, jak i współczynnik regresji są wysoko istotne.

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotność R	0.940444444
R kwadrat	**0.884435752**
Dopasowany R kwadrat	0.884377504
Błąd standardowy	1.412703866
Obserwacje	1986

ANALIZA WARIANCJI

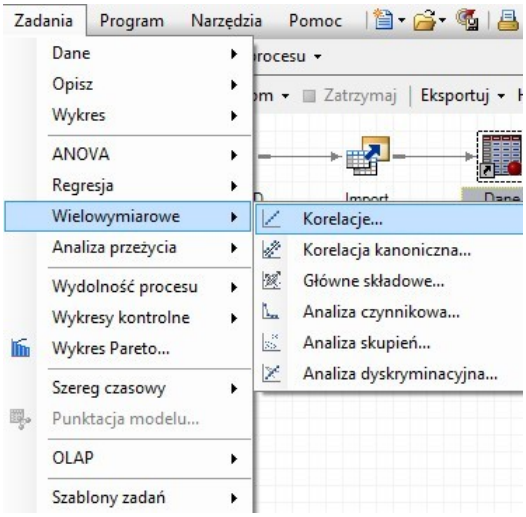
	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	**1**	**30303.07688**	**30303.08**	**15183.94**	**0**
Resztkowy	1984	3959.532711	1.995732		
Razem	1985	34262.60959			

	Współczynniki	*Błąd standardowy*	*t Stat*	***Wartość-p***	*Dolne 95%*	*Górne 95%*
Przecięcie	**67.98809581**	0.118639792	573.0632	**0**	67.75542416	68.22076747
Slonina1	**-0.830978778**	0.006743692	-123.223	**0**	-0.84420424	-0.817753318

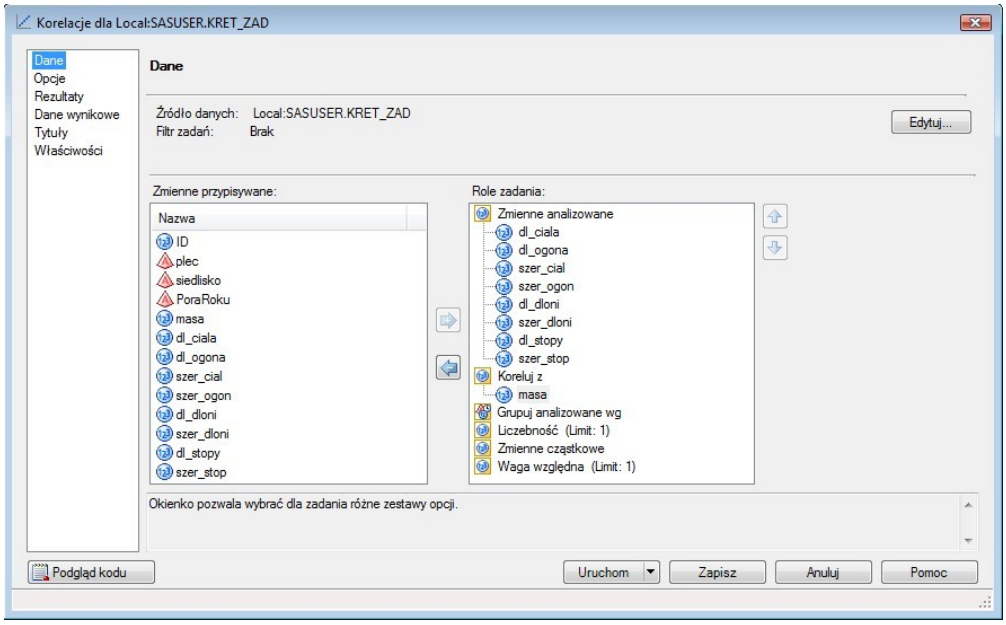
III. Przykład SAS EG:

1. Sprawdź, które z wymiarów ciała są najsilniej związane z masą ciała kretów (**S:\~\bazyXLS\KRET_ZAD.XLS**).

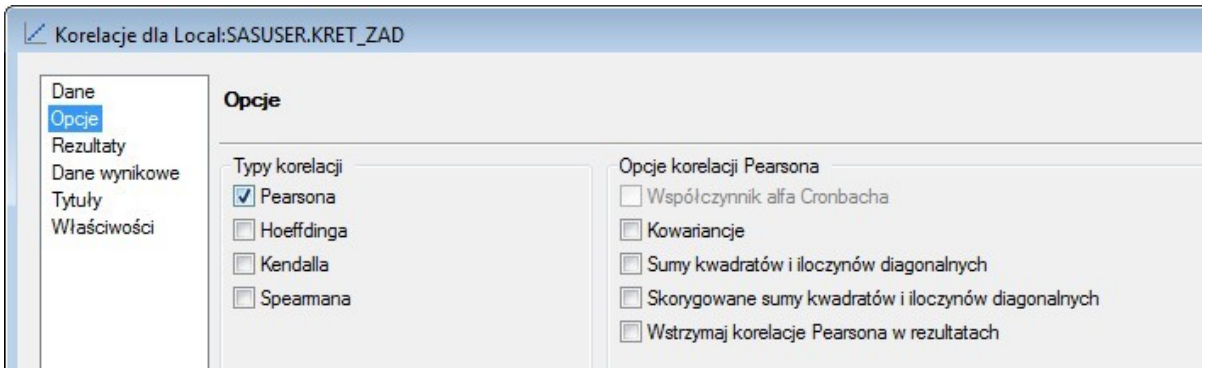
Uruchamiamy kreator korelacji.



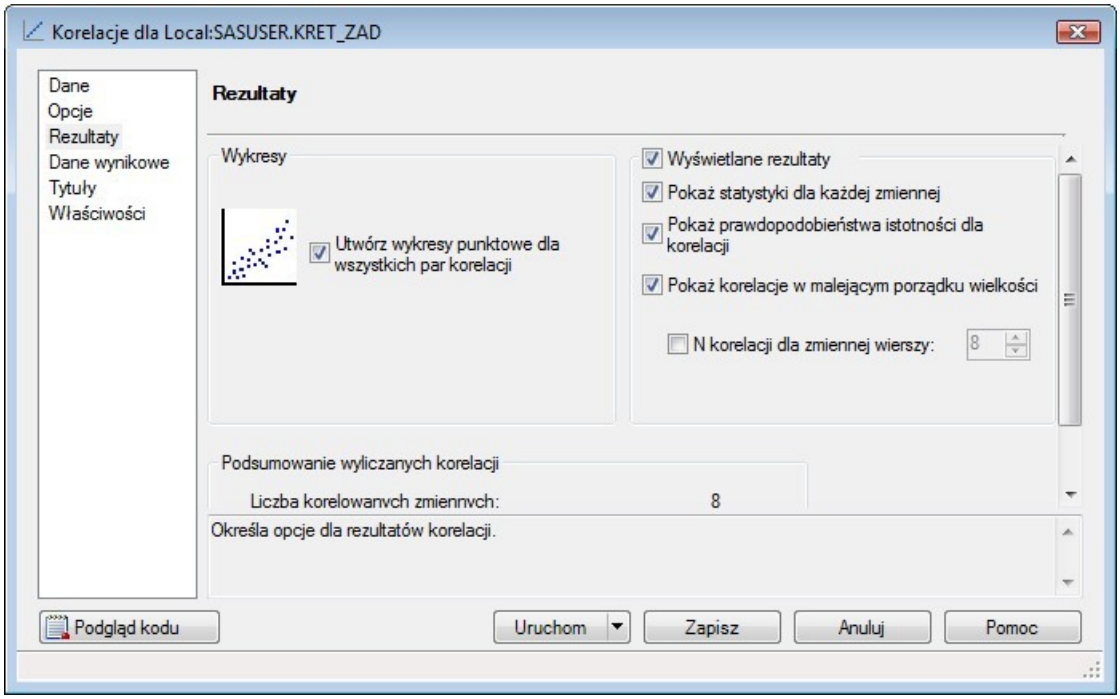
W pierwszym kroku wskazujemy na zmienne, które będziemy korelować z masą ciała oraz samą masę ciała (Koreluj z).



Następnie zaznaczamy pole wyboru **Pearson**, dzięki czemu zostanie obliczony współczynnik korelacji Pearsona.



W kolejnym etapie ustalamy rodzaj statystyki, które chcemy obliczyć. Przy okazji obliczania współczynników zostaną zaprezentowane również podstawowe miary położenia i zmienności. Zostanie również podane prawdopodobieństwo dotyczące współczynnika korelacji. Same zaś współczynniki korelacji zostaną uporządkowane malejąco (*Pokaż korelacje w malejącym porządku wielkości*).

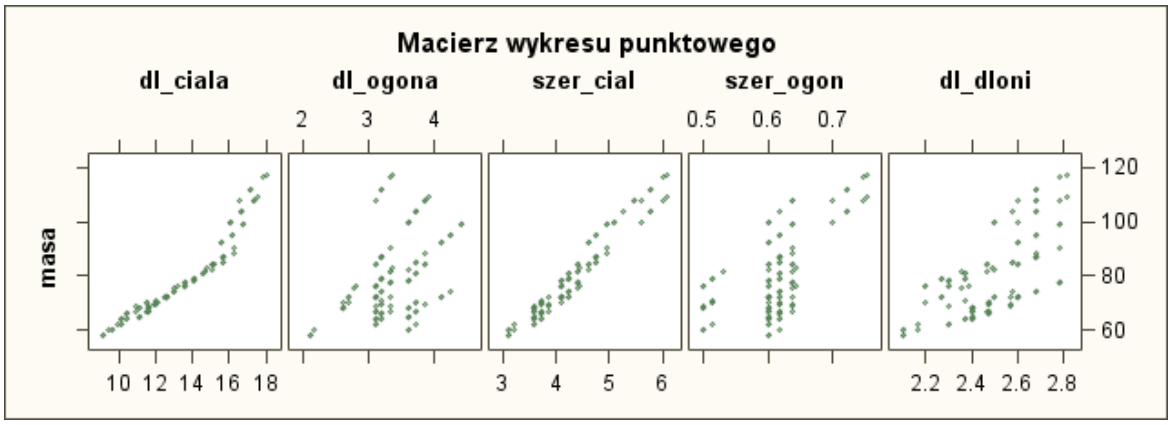


Uzyskane wyniki:

Współczynniki korelacji Pearsona, N = 110
Prawd. > |r| przy H0: Rho=0

	szer_cial	dl_ciala	szer_dloni	dl_dloni	szer_ogon	dl_stopy	szer_stop	dl_ogona
masa	0.97122	0.95378	0.79566	0.67965	0.62321	0.60511	0.51922	0.43058
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Wykresy rozrzutu:

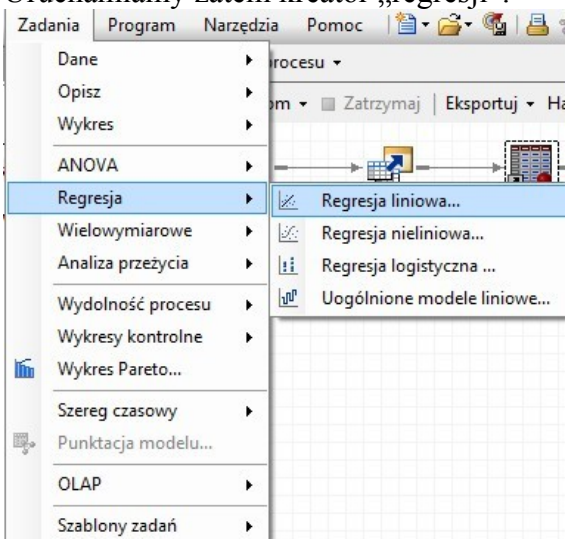


Stwierdzono, że wszystkie obliczone współczynniki korelacji prostoliniowej między masą ciała zwierząt a wymiarami ciała były wysoko istotne. Bardzo wysoką zależność zarejestrowano między długością i szerokością tułowia a masą zwierząt. Wysoką zależność stwierdzono między szerokością i długością dłoni a masą ciała. Pozostałe współczynniki korelacji świadczą o umiarkowanych zależnościach.

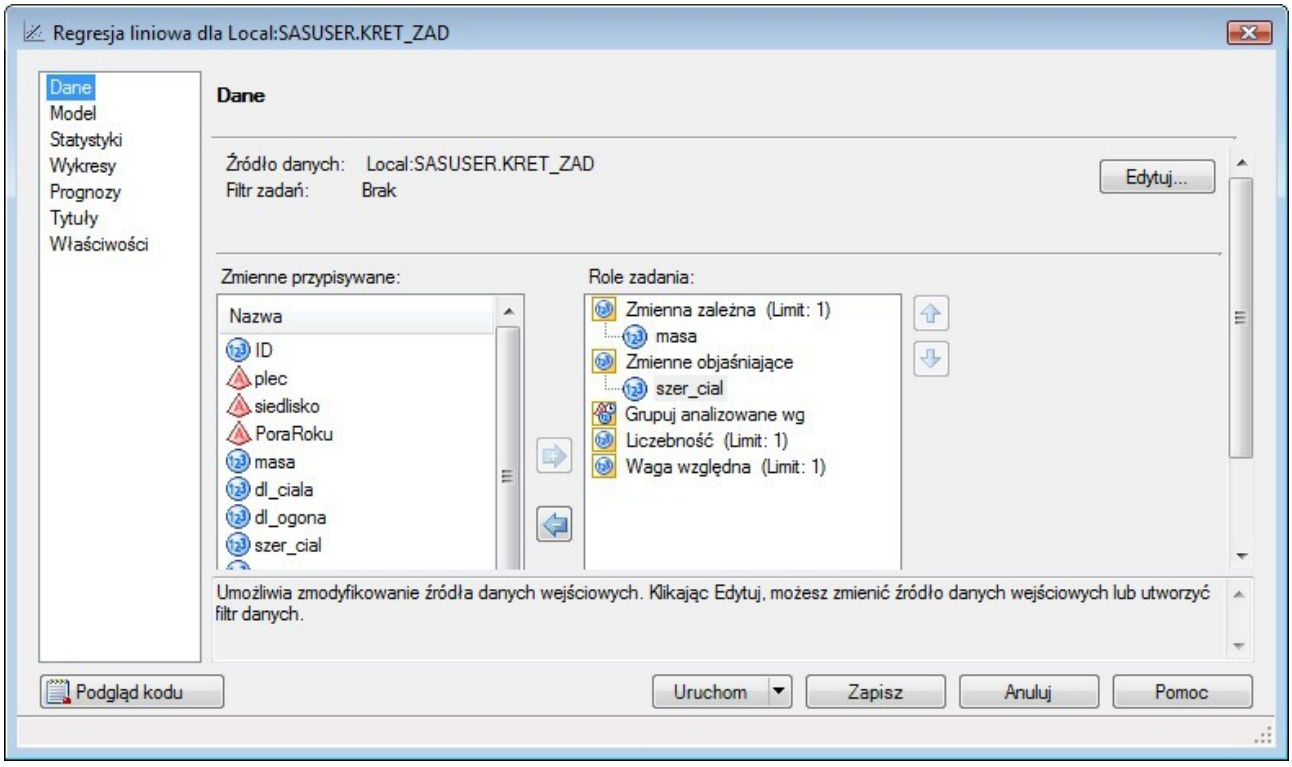
IV. Przykład SAS EG:

Sporządź równanie regresji liniowej, które zostanie zastosowane do przewidywania masy ciała (Zmienna zależna) kretów na podstawie pomiaru szerokości ich ciała (Zmienna objaśniająca).

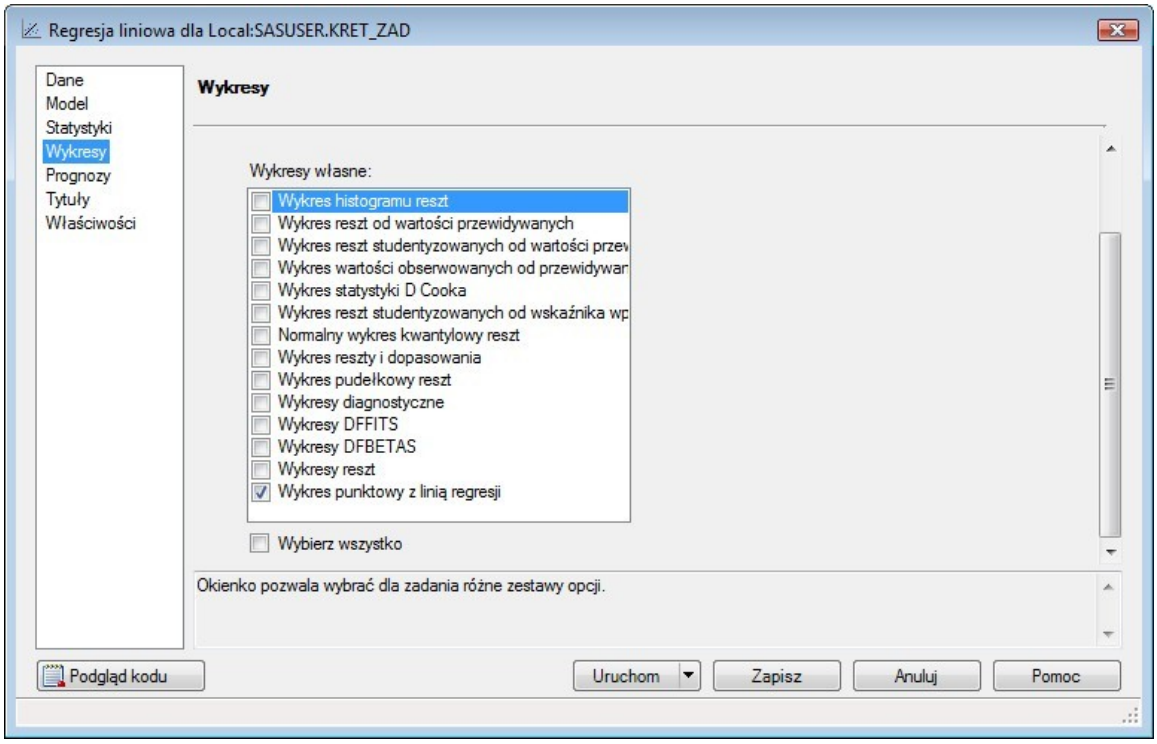
Uruchamiamy zatem kreator „regresji”.



Wskazujemy zmienną zależną oraz objaśniające.



Elementem, na który powinniśmy zwrócić uwagę jest zadanie **Wykres**. Wykonamy tylko jeden wykres, tj. **Wykres punktowy z linią regresji**.



Wyniki:

Model: Linear_Regression_Model

Zmienna zależna: masa

Liczba obserwacji wczytanych	110
Liczba obserwacji użytych	110

Analiza wariancji

Źródło	St. sw.	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	1	21545	21545	1795.70	<.0001
Błąd	108	1295.78409	11.99800		
Razem skorygowane	109	22841			

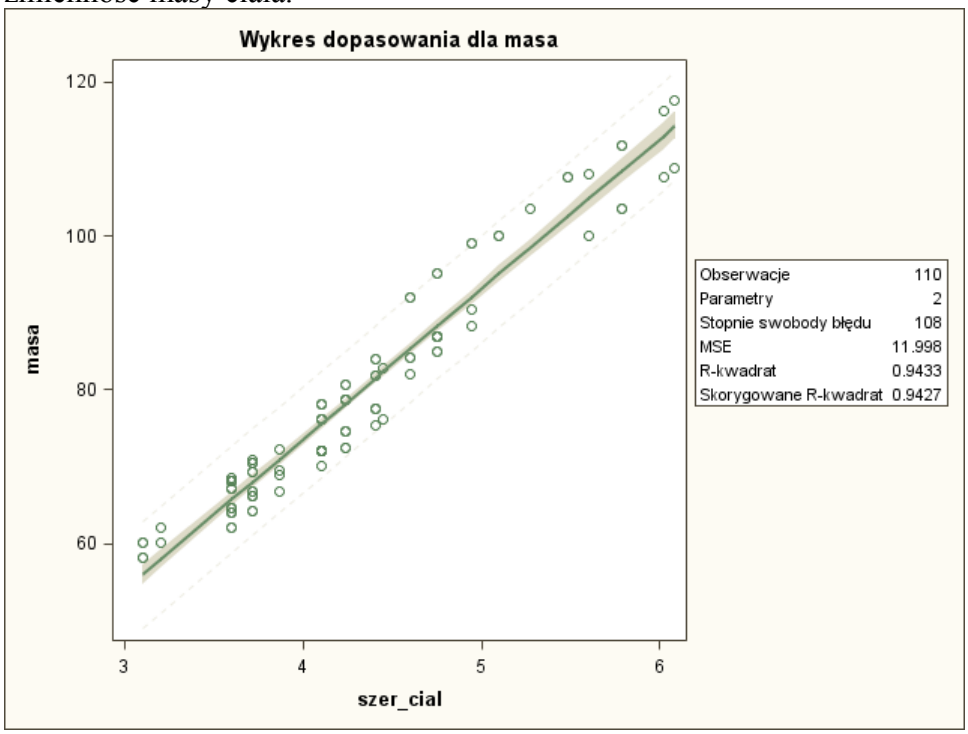
Pierw. bł. śr.-kw.	3.46381	R-kwadrat	0.9433
Średnia zależna	78.89799	Skor. R-kw.	0.9427
Wsp. zmienności	4.39024		

Oceny parametrów

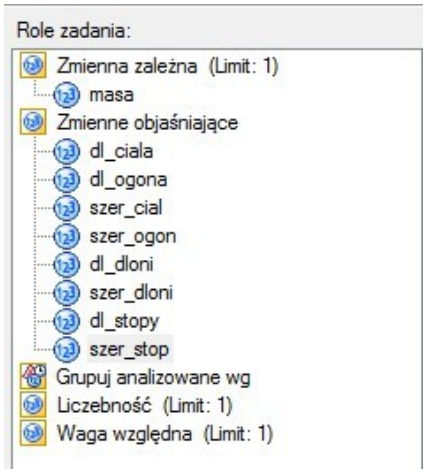
Zmienna	St. sw.	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	1	-4.73446	2.00104	-2.37	0.0198
szer_cial	1	19.56644	0.46174	42.38	<.0001

Legenda
Intercept –wyraz wolny.

Uzyskane wyniki możemy zinterpretować podobnie jak to miało miejsce w przypadku mięsności tuczników. Poniżej umieszczono wykres rozrzutu wraz z miarami jakości modelu. Obliczony współczynnik determinacji świadczy o dobrym dopasowaniu modelu matematyczne do punktów w układzie współrzędnych. Świadczy on o tym, iż model matematyczny dobrze wyjaśnia zmienność masy ciała.



W kolejnym etapie wykonaj równanie regresji wielokrotnej, w ramach której do modelu jako zmienne objaśniające zostaną włączone wszystkie zmienne związane z wymiarami ciała.



Na podstawie uzyskanych wyników możemy wnioskować, że cały model regresji jest istotny – o tym świadczą wyniki analizy wariancji (Pr. > F). Umieszczone w kolumnie **oceny parametrów** wartości pozwalają zbudować model liniowy regresji wielorakiej. Zapisz jego postać. Uwagi wymaga wartość **Skorygowanego współczynnika determinacji**. Jest ona wyższa niż w przypadku modelu z jedną zmienną objaśniającą (0,9427).

Liczba obserwacji wczytanych	110
Liczba obserwacji użytych	110

Analiza wariancji

Źródło	St. sw.	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	8	22168	2771.02634	416.26	<.0001
Błąd	101	672.35759	6.65701		
Razem skorygowane	109	22841			

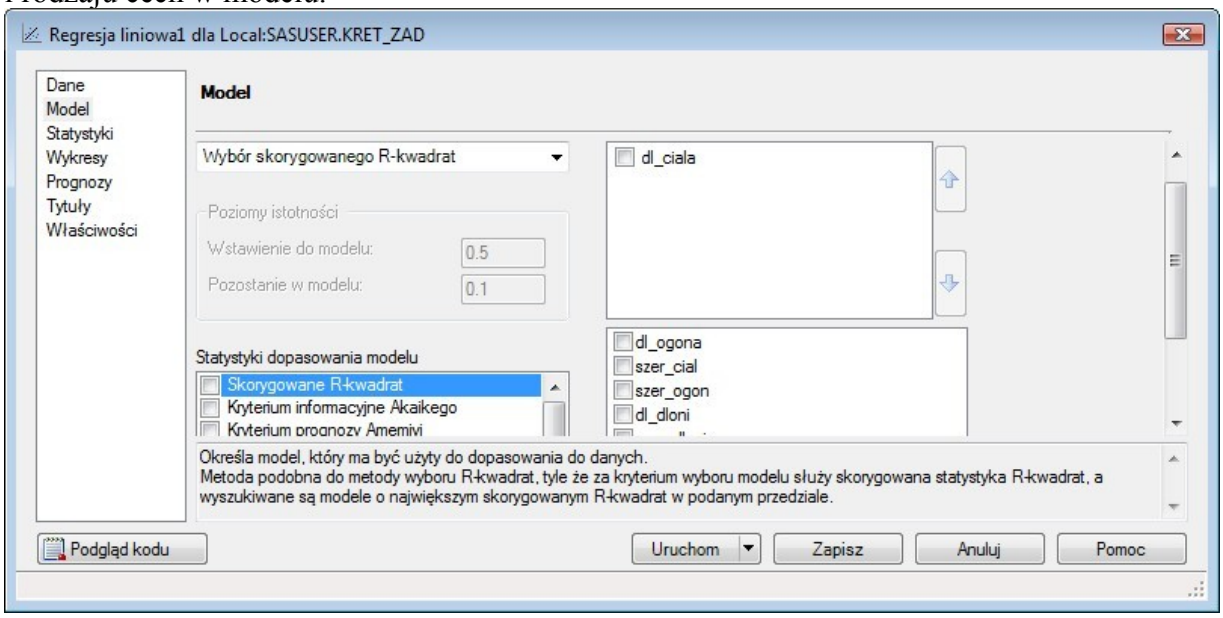
Pierw. bł. śr.-kw.	2.58012	R-kwadrat	0.9706
Średnia zależna	78.89799	Skor. R-kw.	0.9682
Wsp. zmienności	3.27019		

Oceny parametrów

Zmienna	St. sw.	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|
Intercept	1	-7.94166	4.62524	-1.72	0.0890
dl_ciala	1	1.71020	0.40033	4.27	<.0001
dl_ogona	1	0.93574	0.72977	1.28	0.2027
szer_cial	1	13.16614	1.34149	9.81	<.0001
szer_ogon	1	15.20543	9.60115	1.58	0.1164
dl_dloni	1	-8.89518	3.80446	-2.34	0.0214
szer_dloni	1	3.56994	4.39953	0.81	0.4190
dl_stopy	1	6.08728	1.06382	5.72	<.0001
szer_stop	1	-2.78589	1.87209	-1.49	0.1398

V. Przykład SAS EG:

Sprawdź jak kształtują się skorygowane współczynniki determinacji w zależności od liczby i rodzaju cech w modelu.



W tabeli znajdują się obliczone kryteria oceny jakości modelu, tj. R^2 oraz skorygowany współczynnik R^2 w zależności od liczy i rodzaju cech ujętych w modelu. Uzyskane zestawienie pozwala wskazać model najlepiej wyjaśniający zmienność masy ciała.

Liczba wu	r-kwadrat skorygowany	r-kwadrat	Zmienne w Modelu
1	0.9089	0.9097	dl_ciala
7	0.9683	0.9704	dl_ogona szer_cial szer_ogon dl_dloni dl_stopy szer_stop
4	0.9683	0.9694	szer_cial dl_dloni dl_stopy
8	0.9682	0.9706	dl_ogona szer_cial szer_ogon dl_dloni szer_dloni dl_stopy szer_stop
5	0.9682	0.9697	szer_cial szer_ogon dl_dloni dl_stopy
6	0.9681	0.9699	dl_ogona szer_cial szer_ogon dl_dloni dl_stopy
5	0.9681	0.9696	dl_ogona szer_cial dl_dloni dl_stopy
6	0.9681	0.9699	szer_cial szer_ogon dl_dloni dl_stopy szer_stop
6	0.9681	0.9698	dl_ogona szer_cial dl_dloni dl_stopy szer_stop
5	0.9681	0.9695	szer_cial dl_dloni dl_stopy szer_stop
7	0.9680	0.9701	szer_cial szer_ogon dl_dloni szer_dloni dl_stopy szer_stop
5	0.9680	0.9694	szer_cial dl_dloni szer_dloni dl_stopy
6	0.9679	0.9697	szer_cial szer_ogon dl_dloni szer_dloni dl_stopy
7	0.9679	0.9699	dl_ogona szer_cial szer_ogon dl_dloni szer_dloni dl_stopy
6	0.9678	0.9696	dl_ogona szer_cial dl_dloni szer_dloni dl_stopy
6	0.9678	0.9695	szer_cial dl_dloni szer_dloni dl_stopy szer_stop
7	0.9678	0.9698	dl_ogona szer_cial dl_dloni szer_dloni dl_stopy szer_stop
4	0.9675	0.9687	szer_cial szer_dloni dl_stopy
5	0.9674	0.9689	dl_ogona szer_cial szer_dloni dl_stopy

Zadania do samodzielnego wykonania (MS EXCEL / SAS EG).

1. Oblicz współczynniki korelacji Pearsona między temperaturą, pH wody a stężeniem jonów (**S:\~\bazyXLS\ciekiWodne.xls**). Sporządź równanie regresji w odniesieniu do temperatury wody (zmienna niezależna) oraz najsilniej z nią skorelowaną zmienną (zmienna zależna). W tym celu zastosuj narzędzie **Analiza danych** => **Korelacja** oraz **Regresja**.
2. Zbadaj zależność (oblicz współczynniki korelacji) między różnymi rodzajami drobnoustrojów w ściekach o różnym pochodzeniu (**bakterie_ZAD_KOR.xls**).
3. Jakiego rodzaju zależność istnieje między stężeniem pyłu a dwutlenku siarki (**babulice100.xls**)?
4. Oblicz współczynniki korelacji między masą ciała noworodków a długością ich ciała, wynikami skali APGAR oraz wiekiem ich rodziców (**dziecko100.xls**).
5. Wykonaj analizę zależności w odniesieniu do masy ciała oraz wymiarów muszli ślimaka winniczka. Wykonaj analizę regresji wielokrotnej, która pozwoli prognozować masę ciała na podstawie wymiarów muszli (**winniczek.xls**).