Korelacja i regresja liniowa

Martyna Kobielnik

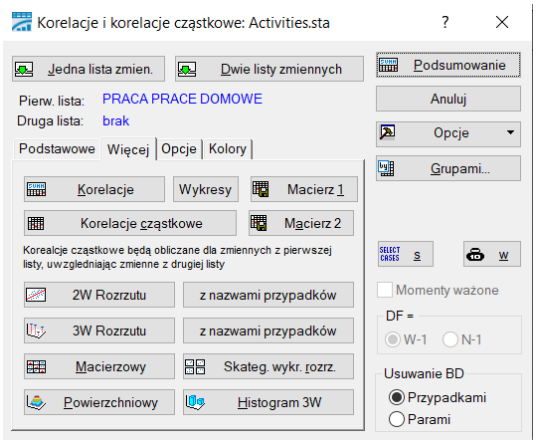
Spis treści

1 Przykłady

Przykład wykonania analizy regresji w programie Statistica zostanie zaprezentowany na przykładzie wybranych zmiennych z arkusza Activities. Warunkiem, umożliwiającym przeprowadzenie wersji parametrycznej jest to, aby zmienne były mierzalne. W przeciwnym wypadku należy skorzystać z modułu *Nieparametryczne*.

W arkuszu znajdują się informacje dotyczące czasu poświęcanego na pewne czynności w pewnych rejonach z podziałem na płeć. Naszym celem będzie sprawdzenie, czy istnieje relacja liniowa łącząca czas poświęcany na pracę z czasem poświęcanym na zadania domowe. Jako zmienną objaśnianą (zależną) przyjmiemy drugą z nich.

W zakładce *Statystyka* wybieramy *Statystyki podstawowe* a następnie *Macierze korelacji*. W otrzymanym oknie (rys. 1a) wybieramy listę zmiennych do analizy. Po kliknięciu na *Korelacje* otrzymamy arkusz widoczny na rys. 1b, który zawiera współczynniki korelacji.



(a) Okno analizy

Korelacje (Activities.sta)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=28 (Braki danych usuwano przypadkami)

Zmienna	Średnia	Odch.std	PRACA	PRACE DOMOWE
PRACA	448,8571	226,9764	1,000000	-0,906398
PRACE DOMOWE	276,9643	198,6067	-0,906398	1,000000

(b) Współczynniki korelacji liniowej

Rysunek 1: Korelacje

Na czerwono oznaczone są te, które przy pomocy odpowiedniego testu zostały ocenione jako istotnie większe od zera (próg dla podświetlenia można dostosować).

Wartość bezwzględną współczynnika korelacji liniowej Pearsona można zinterpretować następująco:

- $|r_{xy}| = 0$ – brak korelacji liniowej
- $0 < |r_{xy}| < 0.1$ – korelacja nikła
- $0.1 \leq |r_{xy}| < 0.3$ – korelacja słaba
- $0.3 \leq |r_{xy}| < 0.5$ – korelacja przeciętna
- $0.5 \leq |r_{xy}| < 0.7$ – korelacja silna
- $0.7 \leq |r_{xy}| < 0.9$ – korelacja bardzo silna
- $0.9 \leq |r_{xy}| < 1$ – korelacja prawie pełna
- $|r_{xy}| = 1$ – korelacja pełna

Ostatni przypadek możliwy jest, jeśli zmienne są ze sobą związane zależnością czysto funkcyjną, np. X określa czas trwania połączenia telefonicznego a Y jest kosztem rozmowy.

Należy mieć na uwadze, że nawet silna korelacja nie oznacza związku przyczynowo–skutkowego. Jest to bardzo często popełniany błąd. Stąd, po przeprowadzeniu analizy, jeśli korelacja została wykryta, przed wyciągnięciem ostatecznych wniosków należy dokładnie zbadać relację między zmiennymi. Przykłady obrazujące pułapki silnej korelacji znaleźć można np. na stronie internetowej `http://tylervigen.com`

W zakładce *Opcje* możemy wybrać rodzaj informacji wyświetlanych w arkuszu wyników. Przykładowy arkusz po zaznaczeniu opcji *Wyświetl dokładną tabelę wyników* przedstawiony został na rys. 2a. Możemy z niego odczytać współczynnik korelacji liniowej, współczynnik determinacji, wartość statystyki testowej wraz z wartością p dla testu istotności współczynnika korelacji oraz parametry α i β prostej regresji. Przykładowo, traktując zmienną PRACA jako zmienną zależną, z drugiej sekcji arkusza odczytujemy współczynniki z kolumn Stała zal: X i Nachylenie zal: X otrzymując:

$$PRACA = -1.04 \cdot PRACE\ DOMOWE + 735.76.$$

Jeśli wybrane zostaną dwie listy, to zestawione ze sobą zostaną każda zmienna z listy pierwszej z każdą zmienną z listy drugiej. Przykładowy arkusz dla kilku zmiennych znajduje się na rys. 2b.

Korelacje (Activities.sta)
Oznaczone wsp. korelacji są istotne z p < ,05000
(Braki danych usuwano przypadkami)

Zmn. X & Zmn. Y	Średnia	Odch.st.	r(X,Y)	r2	t	p	Ważnych	Stała zal: Y	Nachyle zal: Y	Stała zal: X	Nachyle zal: X
PRACA	448,8571	226,9764									
PRACA	448,8571	226,9764	1,000000	1,000000			28	0,0000	1,00000	0,0000	1,00000
PRACA	448,8571	226,9764									
PRACE DOMOWE	276,9643	198,6067	-0,906398	0,821556	-10,9409	0,000000	28	632,9562	-0,79311	735,7563	-1,03587
PRACE DOMOWE	276,9643	198,6067									
PRACA	448,8571	226,9764	-0,906398	0,821556	-10,9409	0,000000	28	735,7563	-1,03587	632,9562	-0,79311
PRACE DOMOWE	276,9643	198,6067									
PRACE DOMOWE	276,9643	198,6067	1,000000	1,000000			28	0,0000	1,00000	0,0000	1,00000

(a) Dokładna tabela wyników

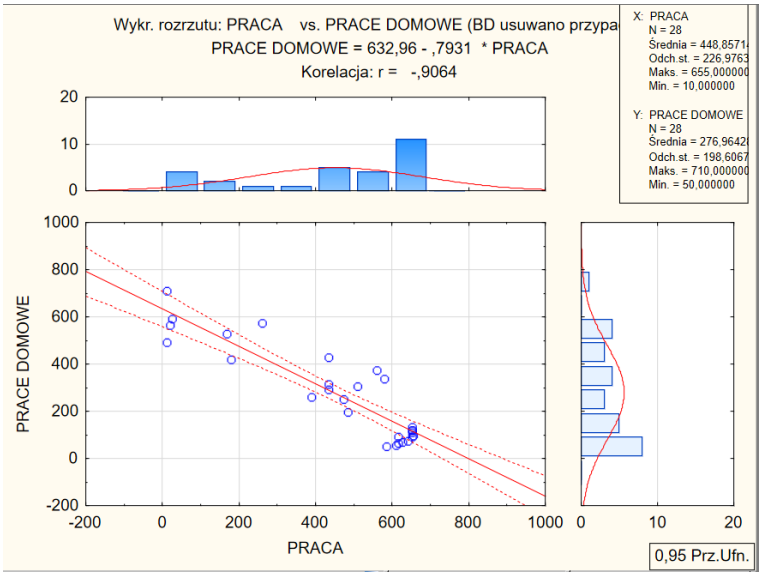
Korelacje (Activities.sta)
Oznaczone wsp. korelacji są istotne z p < ,05000
(Braki danych usuwano przypadkami)

Zmn. X & Zmn. Y	Średnia	Odch.st.	r(X,Y)	r2	t	p	Ważnych	Stała zal: Y	Nachyle zal: Y	Stała zal: X	Nachyle zal: X
PRACA	448,8571	226,9764									
PRACE DOMOWE	276,9643	198,6067	-0,906398	0,821556	-10,9409	0,000000	28	632,9562	-0,79311	735,7563	-1,03587
PRACA	448,8571	226,9764									
ZAKUPY	108,6786	32,5144	-0,654015	0,427736	-4,4084	0,000160	28	150,7311	-0,09369	945,0336	-4,56554
TV	99,4286	39,4090									
PRACE DOMOWE	276,9643	198,6067	-0,205751	0,042334	-1,0721	0,293541	28	380,0627	-1,03691	110,7361	-0,04083
TV	99,4286	39,4090									
ZAKUPY	108,6786	32,5144	0,218571	0,047773	1,1421	0,263816	28	90,7484	0,18033	70,6376	0,26492

(b) Dokładna tabela wyników dla dwóch list

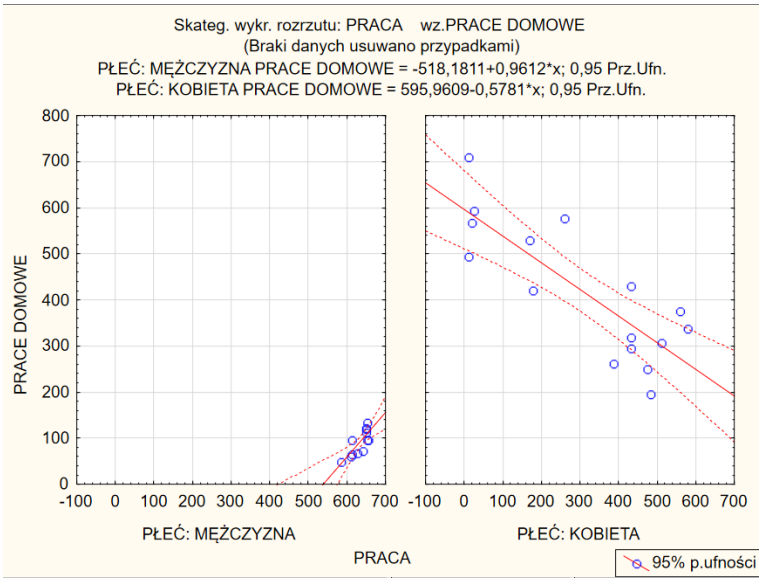
Rysunek 2: Wyniki analizy

Z okna analizy mamy również szybki dostęp do wykresów. Na rys. 3 znajduje się wykres utworzony po wybraniu *Wykresy*. Przedstawia on zestawienie histogramów każdej z dwóch wybranych zmiennych oraz wykres rozrzutu. Wykres rozrzutu posiada dodatkowo prostą regresji wraz z jej przedziałem ufności.



Rysunek 3: Wykres rozrzutu i histogramy zmiennych

Aby przeanalizować zależność dla grup (o ile dostępna jest zmienna grupująca), możemy wykorzystać *Skategoryzowane wykresy rozrzutu*. Na rys. 4 znajdują się wykresy rozrzutu utworzone dla obu płci. Warto zwrócić uwagę na fakt, że dla wszystkich wartości otrzymaliśmy ujemny współczynnik korelacji, jednak wykres dla grupy *MĘŻCZYZNA* ma inne nachylenie niż wykres uwzględniający wszystkie pomiary i wykres dla grupy *KOBIETA*.



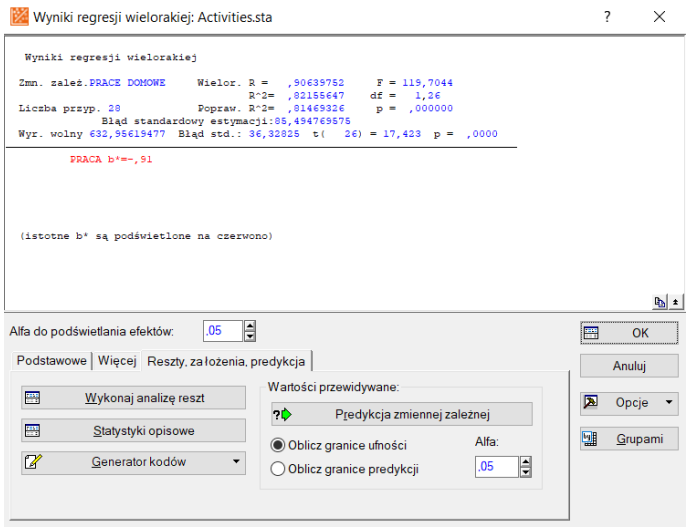
Rysunek 4: Wykresy rozrzutu z podziałem na płeć

W badaniu korelacji możliwe jest zaobserwowanie tzw. paradoksu Simpsona, który polega na tym, że kierunek korelacji dla każdej z grup jest przeciwny do kierunku korelacji, kiedy uwzględniamy wszystkie grupy jednocześnie. Stąd, jeśli dostępna jest zmienna grupująca, warto przeprowadzić dodatkową analizę w grupach, aby lepiej poznać badany zbiór.

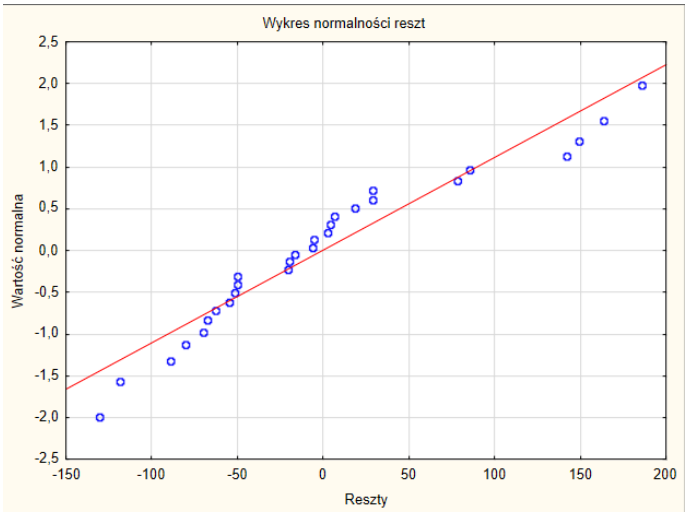
Po wyznaczeniu modelu regresji należy ocenić jego jakość. Żeby był on użyteczny, muszą zostać spełnione następujące warunki:

- Losowość odchyleń – punkty empiryczne powinny być losowo rozmieszczone wokół prostej regresji, nie powinny być możliwe do zaobserwowania żadne prawidłowości w wartościach reszt (np. większe wartości reszt niższego zakresu wartości zmiennej niezależnej)
- Normalność reszt
- Nieobciążoność reszt – reszty powinny mieć wartość średnią równą 0
- Symetria reszt – reszty dodatnie i reszty ujemne powinny mieć mniej więcej takie same liczebności

Aby wykonać analizę reszt, musimy skorzystać z modułu *Regresja wieloraka*. Po wybraniu zmiennych do analizy i przejściu do okna wyników, wybieramy zakładkę *Reszty, założenia, predykcje* widoczną na rys. 5. Spośród dostępnych opcji wybieramy *Analiza reszt*.



Rysunek 5: Wykres normalności reszt



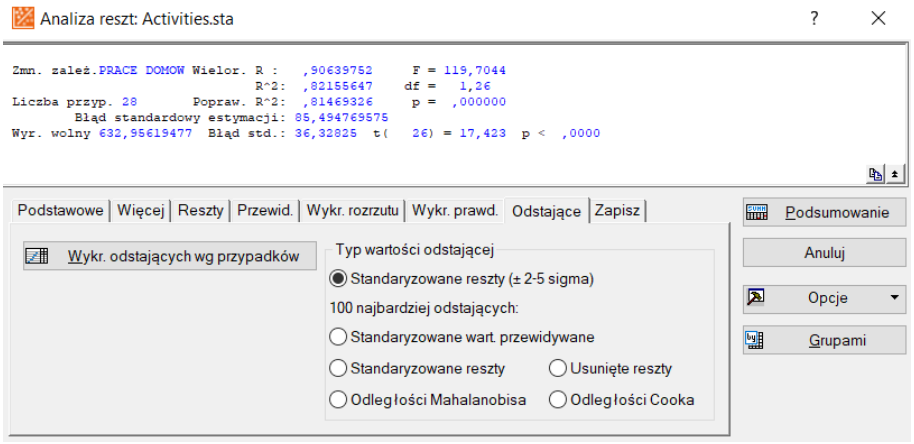
Rysunek 6: Wykres normalności reszt

Na rys. 6 znajduje się wykres normalności reszt, na podstawie którego możemy ocenić, czy spełnione jest odpowiednie założenie. Po wyborze *Podsumowanie: Reszty i przewidywane* możemy zobaczyć arkusz widoczny na rys. 7. Klikając prawym przyciskiem myszy na ten wynik na liście po lewej stronie skoroszytu mamy możliwość wyboru tego arkusza jako aktywny arkusz wejściowy. Dzięki temu możemy obliczyć wartość średnią reszt oraz wykres normalności uzupełniony o wynik testu SW.

	Wartości przewidywane i reszty PRACE DOMOWE								
	Obserw. Wartość	Przewidyw. Wartość	Reszta	Standard. Przewid.	Standard. Reszta	Bł. std. W.przew.	Mahaln. Odległ.	Usunięte Reszta	Cooka Odległ.
EMU	60,000000	149,160690	-89,160690	-0,709954	-1,042879	19,937384	0,504035	-94,288300	0,033072
EWU	250,000000	256,230194	-6,230194	-0,115179	-0,072872	16,267752	0,013266	-6,464236	0,000103
UWU	495,000000	625,025146	-130,025146	1,933493	-1,520855	35,680454	3,738394	-157,448471	0,295358
MMU	65,000000	145,195160	-80,195160	-0,731983	-0,938012	20,151880	0,535799	-84,912804	0,027402
MWU	421,000000	490,989990	-69,989990	1,188922	-0,818646	25,371563	1,413535	-76,749084	0,035486
SMU	50,000000	168,988388	-118,988388	-0,599811	-1,391762	18,932644	0,359773	-125,124390	0,052519
SWU	196,000000	250,678436	-54,678436	-0,146019	-0,639553	16,334642	0,021322	-56,750038	0,008042
EMW	95,000000	115,850189	-20,850189	-0,894996	-0,243877	21,860863	0,801017	-22,308773	0,002226
EWW	307,000000	228,471436	78,528564	-0,269380	0,918519	16,753899	0,072565	81,664642	0,017519
UWW	567,000000	617,094055	-50,094055	1,889435	-0,585931	35,035671	3,569965	-60,204510	0,041638
MMW	97,000000	113,470863	-16,470863	-0,908213	-0,192653	22,007938	0,824851	-17,639750	0,001410
MWW	529,000000	499,714142	29,285858	1,237385	0,342546	25,991322	1,531121	32,268162	0,006583
SMW	72,000000	123,781265	-51,781265	-0,850938	-0,605666	21,379274	0,724096	-55,235275	0,013051

Rysunek 7: Wyniki analizy reszt

Wartości odstające możemy wykryć przechodząc do zakładki *Odstające* widocznej na rys. 8a. Na rys. 8b widoczny jest arkusz zawierający informacje o pomiarach, które zostały na podstawie wybranego kryterium uznane za wartości odstające.



(a) Zakładka *Odstające*

Reszty standaryz.: PRACE DOMOWE (Activities.sta)
Odstające

Nazwa przypadku	Reszty standaryz. −5. −4. −3. ±2. 3. 4. 5.	Obserw. Wartość	Przewidyw. Wartość	Reszta	Standard Przewid.	Standard Reszta	Bł. std. W.przew.
EWY		375,0000	188,8161	186,1839	-0,489667	2,177723	18,05434
Minimum		375,0000	188,8161	186,1839	-0,489667	2,177723	18,05434
Maksim.		375,0000	188,8161	186,1839	-0,489667	2,177723	18,05434
Średnia		375,0000	188,8161	186,1839	-0,489667	2,177723	18,05434
Mediana		375,0000	188,8161	186,1839	-0,489667	2,177723	18,05434

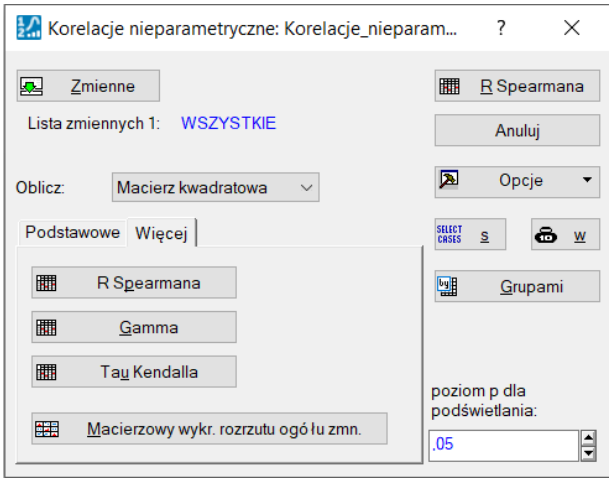
(b) Wykres wartości odstających

Rysunek 8: Identyfikacja wartości odstających

Jeśli dane, które mamy do dyspozycji nie są mierzalne, jak w arkuszu załączonym na PZE, ale dają się uszeregować w postaci rang, możemy wykorzystać statystyki nieparametryczne.

W arkuszu znajdują się wyniki uzyskane przez 10 osób w dwóch konkursach. Posiadamy tylko informacje o tym, które miejsce dana osoba zajęła, a więc po prostu rangi osób w obu konkursach. Chcemy sprawdzić, czy występuje korelacja pomiędzy wynikami.

W module *Nieparametryczne* wybieramy *Korelacje*. W oknie 9 wybieramy analizowane zmienne, a następnie wybieramy rodzaj współczynnika do policzenia.



Rysunek 9: Wybór korelacji nieparametrycznych

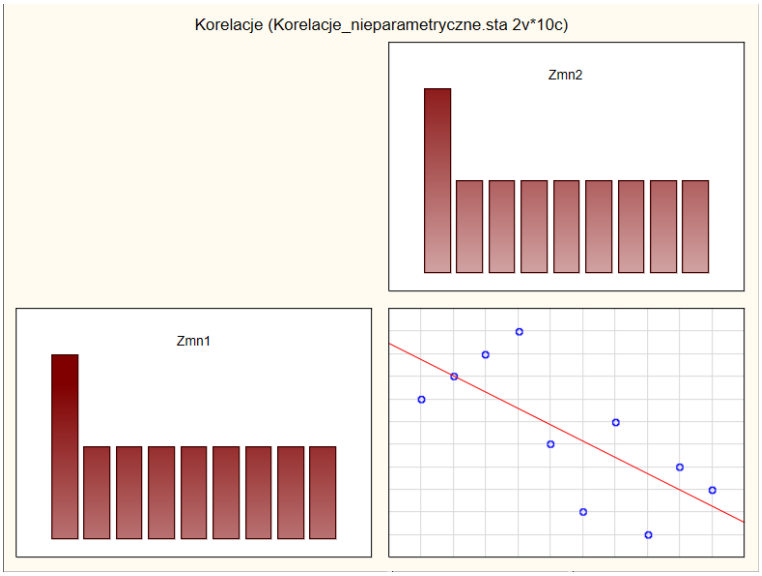
Do wyboru mamy:

- współczynnik R Spearmana, który jest nieparametrycznym odpowiednikiem współczynnika Pearsona obliczonym przy pomocy rang
- współczynnik τ Kendalla, który wyznaczany jest na podstawie prawdopodobieństw
- współczynnik γ, który podobnie jak współczynnik τ wyznaczany jest dla prawdopodobieństw, jednak w przypadku wielu rang wiązanych jest lepszym wyborem.

Na rys. 10 znajduje się arkusz wyników dla współczynnika R Spearmana, a na rys. 11 macierzowy wykres rozrzutu.

	Korelacja porządku rang Spearmana (Korelacje_nieparametryczne. BD usuwane parami Oznaczone wsp. korelacji są istotne z p <,05000	
Zmienna	**Zmn1**	Zmn2
Zmn1	1,000000	-0,721212
Zmn2	-0,721212	1,000000

Rysunek 10: Współczynnik R Spearmana



Rysunek 11: Macierzowy wykres rozrzutu

2 Zadania

1. W arkuszu *School performance* dostępnym w przykładowych arkuszach Statistici znajdź dwie pary zmiennych, dla których korelacja jest istotna statystycznie. Zbuduj dla nich model regresji, utwórz wykres rozrzutu i przeprowadź analizę reszt.