Prognozowanie i symulacje (lab. 4)
Prognozy z wykorzystaniem metod graficznej analizy danych

W programie STATISTICA istnieje możliwość dopasowywania do danych modeli z pewnej grupy funkcji. Do najczęściej stosowanych funkcji może zaliczyć: ***liniową, wielomianową*** *(stopnia 2) oraz* ***wykładniczą.***

Przykład 1 (*Gruźlica w grupach wiekowych)*

Celem analizy będzie prognoza liczby przypadków gruźlicy w Polsce (**zmienna nr 1**) na lata 2025-2027 za pomocą trendu liniowego. Oto kolejne etapy rozwiązania zadania:

- robimy ***WYKRES LINIOWY* (zmiennych)** obrazujący liczbę przypadków gruźlicy w latach 2000-2024 – widać wyraźnie, iż badane zjawisko charakteryzuje się bardzo wyraźną tendencją spadkową o liniowym charakterze;
- wznawiamy analizę (***Ctrl* + *R***) i robimy wykres jeszcze raz, włączając dodatkowo opcję dopasowania trendu liniowego (zakładka ***Więcej* → *Dopasuj* → *Liniowe***);
- w **nagłówku wykresu** podany jest wzór dopasowanej do danych prostej – szczególnie interesujący jest współczynnik przy zmiennej *x*. Proszę na jego podstawie podać o ile w każdym roku, średnio rzecz biorąc, spadała liczba przypadków gruźlicy: ……?
- wzór trendu liniowego posłuży do sporządzenia prognozy na kolejne lata:
  a) kopiujemy wzór prostej regresji z nagłówka wykresu;
  b) minimalizujemy wykres i na końcu arkusza danych dodajemy dwie nowe zmienne, nazywając je *X* oraz *Prognoza przypadków gruźlicy*;
  c) wartości zmiennej *X* będą to numery kolejnych obserwacji, czyli: 1, 2, …, 25 – wartości te wyznaczamy za pomocą formuły „**=v0**”;
  d) wartości zmiennej *Prognoza…* należy wyliczyć za pomocą formuły wyznaczonej i skopiowanej w punkcie a);
  e) proszę dodać trzy nowe przypadki na końcu arkusza danych i wypełnić wartości zmiennej *X* numerami kolejnych obserwacji (lat) – w tym celu wystarczy nacisnąć klawisz F9;
  f) wartości prognoz dla lat 2025-2027 powinny zostać wyliczone automatycznie – **jeżeli nie, wymuszamy obliczenia klawiszem F9**.
- proszę w analogiczny sposób wykonać jeszcze prognozę dla: liczby przypadków gruźlicy u osób w wieku 20-44 lata oraz 45-64 lata.

	13 X	14 Prognoza liczby przypadków gruźlicy
2000	1	10 964
2001	2	10 657
2002	3	10 351
2003	4	10 044
2004	5	9 737
2005	6	9 430
2006	7	9 123
2007	8	8 816
2008	9	8 509
2009	10	8 203
2010	11	7 896
2011	12	7 589
2012	13	7 282
2013	14	6 975
2014	15	6 668
2015	16	6 362
2016	17	6 055
2017	18	5 748
2018	19	5 441
2019	20	5 134
2020	21	4 827
2021	22	4 520
2022	23	4 214
2023	24	3 907
2024	25	3 600
2025	26	3 293
2026	27	2 986
2027	28	2 679

Trend liniowy dopasowany do danych

PROGNOZY

Rok	Prognoza liczby przypadków gruźlicy w Polsce		
	Ogółem	**U osób w wieku 20-44 lata**	**U osób w wieku 45-64 lata**
2025			
2026			
2027			

Przykład 2 (*Transport w Polsce 1990-2024 (R)*)

W analogiczny sposób, wykorzystując możliwość dopasowania innych modeli, proszę sporządzić prognozę poziomu wskazanych w tabeli zjawisk transportowych na lata 2025-2028 za pomocą **modelu kwadratowego** (w tym celu wybieramy dopasowanie: **WIELOMIAN)**.
Uwaga: stopień wielomianu można ustalić w zakładce ***Opcje 2*** (domyślnie jest to funkcja kwadratowa, więc w tym zadaniu nie ma potrzeby dokonywania żadnych zmian).

Rok	Prognoza za pomocą modelu funkcji kwadratowej			
	Przewozy pasażerów (koleje) [mln osób]	**Przewozy pasażerów (samochodowe) [mln osób]**	**Linie kolejowe eksploatowane [km]**	**Długość autostrad [km]**
2025	×[1)]	×[1)]		
2026				
2027				
2028				

[1)] Ponieważ w arkuszu znajdują się już dane z 2025 r. nie ma sensu podawać prognozy na ten rok

Prognozowanie i symulacje (lab. 4)
Prognozy z wykorzystaniem metod graficznej analizy danych

Przykład 3 (*Szkolnictwo wyższe w Polsce 1990-2024 (R)*)

Do prognozy *Liczby studentów ogółem* (zmienna 1) w Polsce wykorzystamy *model wykładniczy*. Przyjmiemy perspektywę roku 2000 i skonstruujemy prognozę na rok 2001 i kolejne lata. Celem tego przykładu będzie zwrócenie uwagi na fakt, iż model bardzo dobrze pasujący do danych w przeszłości nie musi dawać wcale dobrych prognoz. Bardzo ważna jest merytoryczna znajomość danego zjawiska i wiedza o tym, czy dany typ trendu może być kontynuowany.

- proszę sporządzić ***Wykres liniowy*** liczby studentów ogółem w latach **1990-2000** wraz z dopasowaniem w postaci funkcji wykładniczej (aby zawęzić wykonanie wykresu do ww. lat należy znaleźć przycisk SELECT CASES Warunki selekcji i wybrać przypadki od 1 do 11);
- funkcja wykładnicza **niemal idealnie** pasuje do danych – proszę wykorzystać jej wzór i wyznaczyć prognozy na rok 2001 i kolejne lata w arkuszu danych – w sposób analogiczny jak w poprzednich przykładach;
- porównajmy wartości prognozowane na lata 2001-2003 z faktyczną liczbą studentów w tych latach (czyli liczbami z 1. kolumny) – jakie decyzje, związane na przykład z rozbudową infrastruktury albo zatrudnianiem wykładowców, podjęłaby uczelnia, która bazowałaby na tych prognozach? A jaka była prognoza liczby studentów na 2023 czy 2024 rok z perspektywy roku 2000, w stosunku do faktycznej liczby studentów w tym roku?

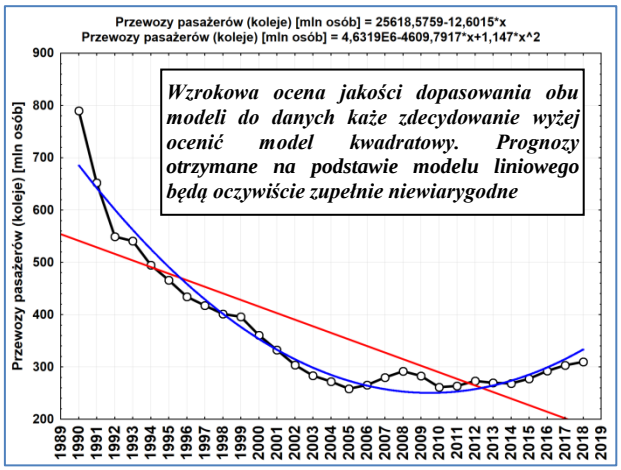
Przykład 4 (*Transport w Polsce 1990-2024 (R)*)

Alternatywnym sposobem prezentacji przebiegu zmienności danych jest ***WYKRES ROZRZUTU***. Zaletą wykresu liniowego jest **łatwość jego wykonania**, **automatyczny opis osi poziomej** (jeśli arkusz jest odpowiednio przygotowany) oraz brak konieczności **tworzenia zmiennej obrazującej współrzędną *X*** w arkuszu, ponieważ program automatycznie numeruje przypadki (począwszy od 1). Jednak czasem warto wykorzystać ***Wykres rozrzutu***, który oferuje pewne dodatkowe możliwości:

- **pomiary zmiennej prognozowanej nie muszą być dokonywane w jednakowych odstępach czasu;**
- **można wyznaczyć pewne elementarne miary dopasowania modelu liniowego do danych.**

Celem analizy jest, podobnie jak w przykładzie 2, sporządzenie prognozy ***Liczby pasażerów przewożonych koleją*** na lata **2026-2028**, Tym razem wykorzystamy w tym celu wykres rozrzutu (***Wykresy* / *Wykresy 2W* / *Wykresy rozrzutu***). Ponieważ wykres rozrzutu wymaga wprowadzenia **dwóch zmiennych**, w arkuszu danych dodamy zmienną ***ROK*** i wypełnimy ją kolejnymi obserwowanymi latami (1990, 1991, …), wykorzystując odpowiednio **zmodyfikowaną** formułę „=v0” (co zrobić, żeby zamiast 1, 2, … pojawiły się wartości 1990, 1991, …?).

W celu utworzenia wykresu rozrzutu na pierwszej liście zmiennych wskazujemy ***ROK*** a na drugiej ***Przewozy pasażerów koleją***. W zakładce ***Więcej*** wskazujemy ***Dopasowanie liniowe*** i wykonujemy wykres. Do gotowego wykresu dodajemy kolejne dopasowanie – tym razem w postaci ***Modelu wielomianowego (stopnia 2)***, czyli funkcję kwadratową. Proszę tak sformatować wykres, by wyglądał jak ten na rysunku obok (oczywiście z danymi do 2025 r.).



W ***nagłówku wykresu*** zamieszczone są wzory obu modeli. Zostaną one wykorzystane do sporządzenia prognoz. Postępujemy w analogiczny sposób jak w poprzednich przykładach – **zmieniamy nazwę zmiennej *ROK* na *X*,** dodajemy 3 przypadki do arkusza danych, dodajemy dwie zmienne (*Prognoza liniowa* i *Prognoza kwadratowa*) oraz wyliczamy ich wartości za pomocą formuł podanych w nagłówku wykresu rozrzutu.

Wykres rozrzutu pozwala uzyskać pewne dodatkowe informacje na temat jakości dopasowania modelu do danych. Niestety, jest to możliwe tylko dla modelu liniowego (dla bardziej złożonych modeli należy wykorzystać analizę regresji, o czym będzie mowa na kolejnych zajęciach). Proszę sporządzić wykres rozrzutu, obrazujący zmienne ***ROK*** i ***Przewozy pasażerów koleją*** z dopasowaniem liniowym, a w zakładce ***Więcej*** wybrać dodatkowo dwie statystyki: ***R kwadrat. (dopasowania liniowego)*** oraz ***Współczynnik korelacji i p (dopasowania liniowego)***. Wartość R^2 wyrażamy w procentach zaś wartość p z dokładnością do 4 miejsc po przecinku.

Proszę podać te wartości: R^2 = ……., p = …….

Oto znaczenie praktyczne tych miar:

***R*^{2}** – to tzw. **współczynnik determinacji**, który określa jakość dopasowania modelu do danych i przyjmuje wartości z zakresu 0-100%.

p – to tzw**. prawdopodobieństwo testowe**, które określa, czy model jest dopasowany do danych w sposób statystycznie istotny (jest tak, jeśli $p<0,05$).

Na razie proszę zapamiętać, że współczynnik determinacji wyrażamy w procentach – i im większa jego wartość tym lepiej model dopasowany do danych (w przeszłości) i potencjalnie lepsze prognozy. Wartość prawdopodobieństwa testowego p powinna być niska ($<0,05$), bowiem tylko wtedy położenie prostej trendu jest wyznaczone z odpowiednią dokładnością – jeśli $p\geq0,05$ wtedy istnieje duże ryzyko, że wyznaczony model nie pasuje do danych, kierunek nachylenia funkcji liniowej jest w dużej mierze losowy.