

# Prognozowanie i symulacje (lab. 4)

## Prognozy z wykorzystaniem metod graficznej analizy danych

### Przykład 1 (Transport w Polsce 1990-2022 (R))

W programie STATISTICA istnieje możliwość dopasowywania do danych modeli z pewnej grupy funkcji. W szczególności wykorzystamy trzy rodzaje prostych, acz użytecznych funkcji: liniową, wielomianową (stopnia 2) oraz wykładniczą.

**UWAGA! Ponieważ dane w latach 2020, 2021 i 2022 były nieporównywalne (lockdowny i ich fatalne konsekwencje dla życia społecznego i gospodarczego) proszę zapisać plik pod nową nazwą, usunąć trzy ostatnie przypadki i wykonywać prognozy z perspektywy roku 2019. Dotyczy to przykładu 1, 2 i 4.**

Analiza dotyczyć będzie liczby zarejestrowanych samochodów ciężarowych, dla której sporządzimy prognozę w oparciu o trend liniowy. Oto kolejne etapy rozwiązania zadania:

- robimy WYKRES LINIOWY (zmiennych) obrazujący zmianę poziomu badanej cechy w latach 1990-2019 – widać, iż badane zjawisko charakteryzuje się bardzo wyraźną tendencją wzrostową 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, wzrastała liczba zarejestrowanych w Polsce samochodów ciężarowych: .....
- wyznaczony wzór trendu liniowego posłuży do sporządzenia prognozy na kolejne lata – w tym celu:
  - 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 liniowa liczby samochodów ciężarowych;
  - c) wartości zmiennej  $X$  będą to numery kolejnych obserwacji, czyli: 1, 2, ..., 30 – 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 cztery 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 2020-2023 powinny zostać wyliczone automatycznie – **jeżeli nie, wymuszamy obliczenia klawiszem F9.**
- proszę w analogiczny sposób wykonać prognozę dla: liczby zarejestrowanych samochodów osobowych, autobusów i liczby pasażerów podróżujących koleją za pomocą modelu trendu liniowego, uzupełnić poniższą tabelę i skomentować (na podstawie wykresów), czy prognozy liniowe można uznać za wiarygodne?

	16 $X$	17 Prognoza liczby ciężarówek [mln]
1990	1	1,016
1991	2	1,102
1992	3	1,188
1993	4	1,273
1994	5	1,359
1995	6	1,445
1996	7	1,531
1997	8	1,617
1998	9	1,702
1999	10	1,788
2000	11	1,874
2001	12	1,960
2002	13	2,046
2003	14	2,131
2004	15	2,217
2005	16	2,303
2006	17	2,389
2007	18	2,475
2008	19	2,560
2009	20	2,646
2010	21	2,732
2011	22	2,818
2012	23	2,904
2013	24	2,989
2014	25	3,075
2015	26	3,161
2016	27	3,247
2017	28	3,333
2018	29	3,418
2019	30	3,504
2020	31	3,590
2021	32	3,676
2022	33	3,762
2023	34	3,847

Trend liniowy  
dopasowany  
do danych

PROGNOZY

Rok	Zarejestrowane samochody osobowe [mln sztuk]	Zarejestrowane autobusy [tys. sztuk]	Pasażerowie przewożeni koleją [mln osób]
2020			
2021			
2022			
2023			

### Przykład 2 (Transport w Polsce 1990-2022 (R))

W analogiczny sposób, wykorzystując możliwość dopasowania innych modeli, proszę sporządzić prognozę na lata 2020-2023 dla poniższych cech za pomocą modelu kwadratowego (w tym celu wybieramy dopasowanie: **WIELOMIAN**).

Uwaga: stopień wielomianu można ustalić w zakładce Opcje 2 (domyślnie jest to funkcja kwadratowa).


Rok	Zarejestrowane samochody osobowe [mln sztuk]	Zarejestrowane autobusy [tys. sztuk]	Pasażerowie przewożeni koleją [mln osób]
2020			
2021			
2022			
2023			

# Prognozowanie i symulacje (lab. 4)

## Prognozy z wykorzystaniem metod graficznej analizy danych

### Przykład 3 (Szkolnictwo wyższe w Polsce 1990-2021 (R))

Do prognozy *Liczyby studentów ogółem* (zmienna 1) w Polsce wykorzystamy *model wykładniczy*. Przyjmujemy perspektywę roku 2000 i konstruujemy 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  **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 2019 czy 2021 rok z perspektywy roku 2000, w stosunku do faktycznej liczby studentów w tym roku?

### Przykład 4 (Transport w Polsce 1990-2022 (R))

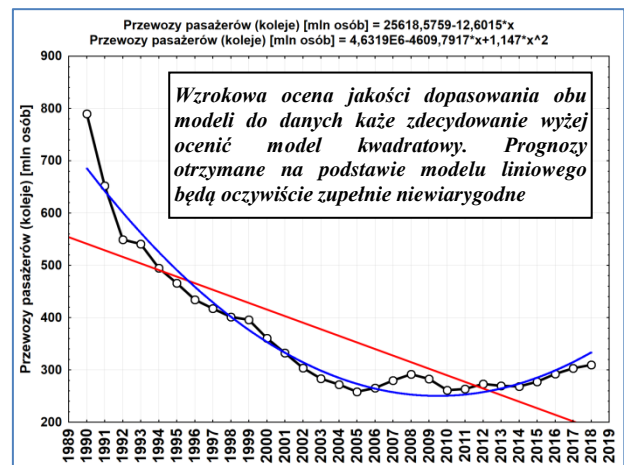
Alternatywnym sposobem prezentacji przebiegu zmienności danych jest wykres rozrzutu. Zaletą wykresu liniowego jest **łatwość jego wykonania, automatyczny opis osi poziomej** (jeśli arkusz jest odpowiednio przygotowany) oraz brak konieczności **tworzenia zmiennej obrazującej współrzędną X** w arkuszu, ponieważ program automatycznie numeruje przypadki (począwszy do 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 sporządzenie prognozy *Liczyby pasażerów przewożonych koleją* na lata 2020-2023, 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 zamieszczony na rysunku.

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 4 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:  $R^2 = \dots\dots\dots$ ,  $p = \dots\dots\dots$

#### 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  $p \geq 0,05$  wtedy istnieje duże ryzyko, że wyznaczony model nie pasuje do danych, kierunek nachylenia funkcji liniowej jest w dużej mierze losowy.