Prognozowanie i symulacje

Wykład 5:
Analiza regresji w prognozowaniu

Co to jest analiza regresji?

Celem tzw. ANALIZY REGRESJI jest badanie związku pomiędzy zmiennymi niezależnymi (objaśniającymi) a zmienną zależną (objaśnianą), która zwykle ma charakter liczbowy.

W naukach społecznych, przyrodniczych (zwłaszcza w medycynie) i ekonomicznych analiza regresji jest szeroko stosowana, jako narzędzie badawcze pozwalające opisać i zrozumieć zjawiska wielowymiarowe. Model regresji może służyć do dokonania prognozy (predykcji) wartości zmiennej zależnej dla „nowych" obiektów (np. kolejnych okresów czasowych).

W klasycznej analizie regresji wielokrotnej model ma postać:

$$Y_i = b_0 + b_1 X_{i1} + \ldots + b_k X_{ik} + e_i \qquad (i = 1, 2, \ldots, n)$$

i pozwala odpowiedzieć na pytanie "jak zmienne niezależne *X* opisują poziom zmiennej zależnej *Y*".

Oznaczenia:

i, n – numer i liczba analizowanych przypadków,
k – liczba zmiennych niezależnych, e_i – błąd modelu dla i-tego przypadku.

Interpretacja parametrów modelu regresji

Etapy analizy regresji są następujące:

- **stawiamy problem badawczy i pozyskujemy odpowiednie dane;**
- **szukamy parametrów modelu regresji, tak by „pasował" on jak najlepiej do posiadanych danych, czyli znajdujemy konkretne wartości parametrów $b_0, b_1, \ldots, b_k$.**

$$Y_i = b_0 + b_1 X_{i1} + \ldots + b_k X_{ik} + e_i$$

- **oceniamy jakość dopasowania modelu do danych;**
- **jeżeli model dobrze odzwierciedla zależności pomiędzy zmiennymi *X* a zmienną *Y* dokonujemy interpretacji jego parametrów.**

Parametr b_0 interpretujemy jako przeciętny (oczekiwany) poziom zmiennej objaśnianej *Y* gdy wszystkie zmienne objaśniające *X* przyjmują wartość 0 (najczęściej jest to nierealna kombinacja, więc interpretację b_0 można wtedy pominąć).

Wzrost wartości zmiennej objaśniającej X_i o jednostkę powoduje zmianę wartości oczekiwanej zmiennej zależnej o b_i jednostek, przy założeniu, że pozostałe zmiennej niezależne zachowują stałe wartości.

Analiza regresji w analizie szeregów czasowych

W przypadku analizy szeregów czasowych, rolę zmiennej objaśniającej pełni zmienna czasowa (we wzorach oznaczana symbolem *t*, w arkuszach danych częściej jako *X*) i/lub jej przekształcenia funkcyjne.

Model trendu liniowego dla szeregu czasowego przyjmuje postać:

$$Y_t = b_0 + b_1 t + e_t \quad (t \text{ – numer okresu czasowego})$$

Parametr b_1 interpretować można jako średnioroczny przyrost prognozowanej wartości w jednostce czasu.

Model funkcji kwadratowej będzie miał postać:

$$Y_t = b_0 + b_1 t + b_2 t^2 + e_t$$

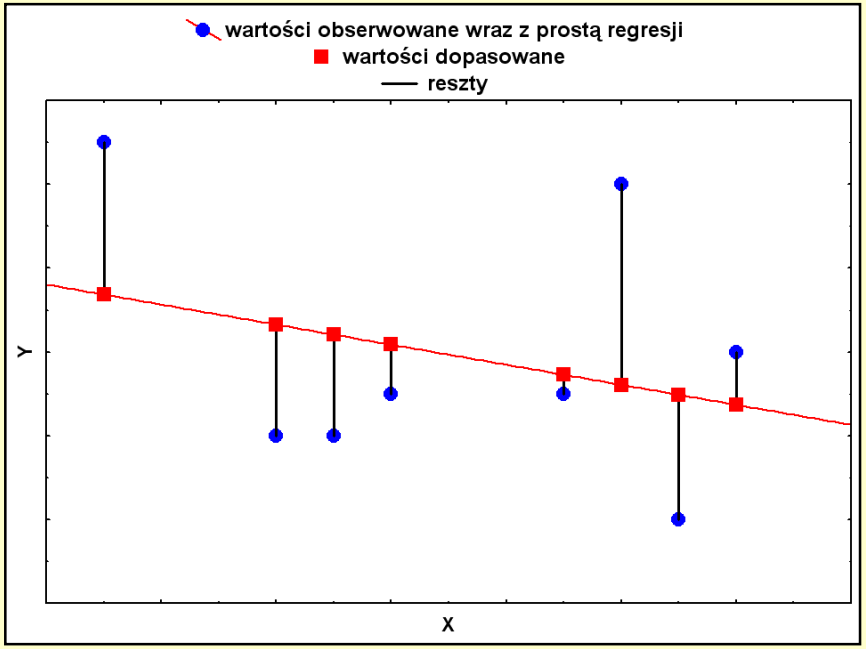
Jest to tak zwany model pozornie nieliniowy, bowiem jeśli podstawimy zamiast t – X_1, a zamiast t^2 – X_2, otrzymamy model liniowy z dwiema zmiennymi niezależnymi.

UWAGA!!! Prognozowanie za pomocą modelu regresji dla danych czasowych wymaga najczęściej dodania w arkuszu danych „sztucznych” zmiennych niezależnych (t, t^2, czy innych przekształceń zmiennej czasowej).

Jak wyznaczany jest model regresji (MNK)?

Wzór modelu regresji jest wyznaczany w taki sposób, by zminimalizować różnicę pomiędzy wartością modelowaną, a faktyczną wartością zmiennej zależnej (*Y*) dla poszczególnych obiektów (w analizie danych czasowych, dla poszczególnych okresów czasowych).

W praktyce, najczęściej przedmiotem optymalizacji jest suma kwadratów odchyleń wartości modelowanych od rzeczywistych pomiarów (tzw. suma kwadratów reszt). Taka metoda dopasowywania modelu do danych nosi nazwę *METODY NAJMNIEJSZYCH KWADRATÓW (MNK).*



Model regresji w programie *STATISTICA*

W programie *STATISTICA* analiza regresji dostępna jest w module *REGRESJA WIELORAKA*.

Możliwość wyznaczenia pewnych modeli liniowych i nieliniowych względem jednej zmiennej niezależnej (a więc na przykład dla szeregów czasowych), udostępniona jest także podczas graficznej analizy danych (za pomocą wykresów liniowych i wykresów rozrzutu). Opis możliwości wykorzystania tych narzędzi do sporządzania prostych prognoz przedstawiono na poprzednim wykładzie.

Moduł *REGRESJA WIELORAKA* pozwala na:

- **wyznaczenie wzoru modelu regresji;**
- **ocenę jego dopasowania do danych;**
- **ocenę istotności poszczególnych zmiennych;**
- **przeprowadzenia tzw. analizy reszt i określenie wpływu na kształt modelu ewentualnych obserwacji odstających;**
- **sporządzenie prognozy punktowej i przedziałowej (z określonym poziomem ufności).**

Czy modele muszą mieć postać liniową?

W programie *STATISTICA* procedura estymacji i weryfikacji modelu liniowego dokonywana jest w module *REGRESJA WIELOKROTNA* (warto wspomnieć, że możliwość oszacowania parametrów modelu regresji i pewnych podstawowych miar jakości jego dopasowania stwarza także arkusz kalkulacyjny *Excel*).

Chociaż natura modelu podlegającego analizie musi być liniowa, to za pomocą formuł arkusza danych bez większych trudności możemy wprowadzać także bardziej skomplikowane typy modeli: np. model kwadratowy, wielomianowy, hiperboliczny (wystarczy w tym celu dodać nową zmienną i nadać jej wartości według interesującej nas formuły).

Bardziej wyrafinowanym narzędziem służącym do konstruowania modeli nieliniowych jest moduł *ESTYMACJI NIELINIOWEJ*, który będzie omawiany na jednym z kolejnych wykładów.

Analiza regresji – przykład wprowadzający (1)

Poniższy przykład dotyczy danych zawartych w pliku ***Efekty rehabilitacji***, zaś celem analizy jest ocena zależności końcowej sprawności chorych (zmienna *Y*) od ***wieku***, ***płci*** i ***wyjściowej sprawności***.

Miara sprawności ma zakres 0-20 pkt.
0 pkt. - sprawność minimalna, 20 pkt. - sprawność maksymalna

1 Wiek X_1	2 Płeć X_2	3 Czy to pierwsza rehabilitacja?	4 Poziom sprawności (przed rehabilitacją) X_3	5 Poziom sprawności (po rehabilitacji) Y
45	kobieta	tak	17	18
61	mężczyzna	tak	2	6
75	mężczyzna	tak	0	5
78	mężczyzna	tak	2	8
84	kobieta	tak	3	11
71	kobieta	tak	5	6
52	mężczyzna	tak	0	12

Aby wykonać analizę, otwieramy w programie *STATISTICA* odpowiedni plik i w menu ***STATYSTYKA*** znajdujemy analizę ***REGRESJA WIELORAKA***.

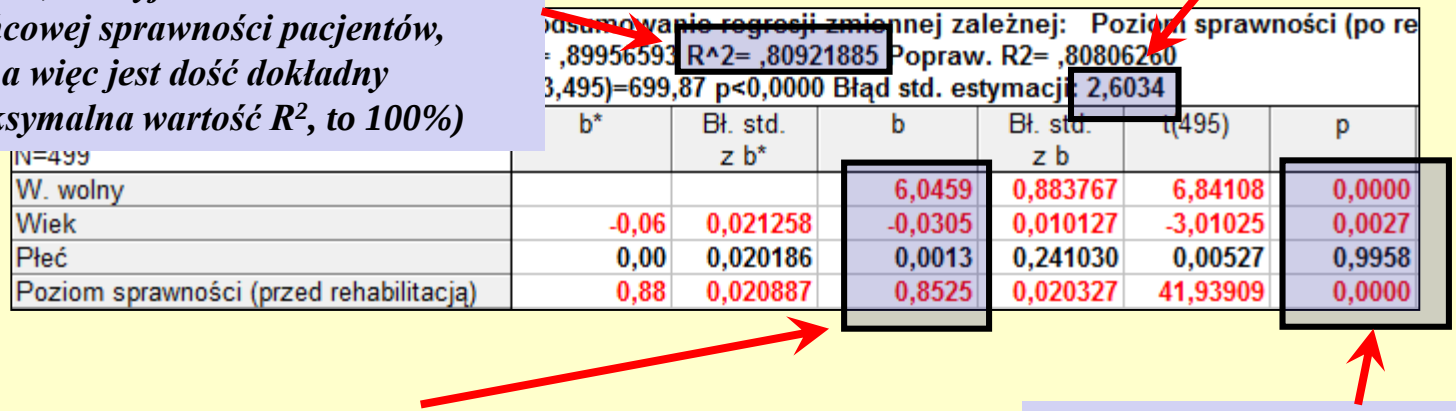
Na liście zmiennych zależnych wskazujemy ***Poziom sprawności po rehabilitacji***, zaś jako zmienne niezależne wybieramy ***wiek***, ***płeć*** i ***poziom sprawności przed rehabilitacją***.

Przykład wprowadzający – wstępne wyniki (2)

Po dokonaniu wyboru zmiennych i zatwierdzeniu przyciskiem *OK*., przechodzimy do okna *Wyniki regresji wielorakiej*, gdzie w zakładce *Podstawowe* wywołujemy wyniki przyciskiem *Podsumowanie: wyniki regresji*.

Wartość współczynnika determinacji R^2, podajemy w procentach. Poniższy model w 80,9% wyjaśnia zmienność końcowej sprawności pacjentów, a więc jest dość dokładny (maksymalna wartość R^2, to 100%)

Błąd standardowy estymacji pozwala stwierdzić, iż model przybliża końcową sprawność pacjentów z dokładnością ±2,60 pkt



,89956593 R^2= ,80921885 Popraw. R2= ,80806260
3,495)=699,87 p<0,0000 Błąd std. estymacji: 2,6034

N=499	b*	Bł. std. z b*	b	Bł. std. z b	t(495)	p
W. wolny			6,0459	0,883767	6,84108	0,0000
Wiek	-0,06	0,021258	-0,0305	0,010127	-3,01025	0,0027
Płeć	0,00	0,020186	0,0013	0,241030	0,00527	0,9958
Poziom sprawności (przed rehabilitacją)	0,88	0,020887	0,8525	0,020327	41,93909	0,0000

W kolumnie „B" podane są wartości współczynników modelu, który, jak widać, przyjmuje postać:

Końcowy poziom sprawności = 6,0459 – 0,0305 · Wiek + + 0,0013 · Płeć + 0,8525 · Wyjściowy poziom sprawności

Ale, jak wynika z komentarza po prawej stronie, z tego modelu należy usunąć zmienną PŁEĆ i oszacować go ponownie.

Wartości prawdopodobieństwa testowego p pozwalają na stwierdzenie, iż poza zmienną płeć, pozostałe czynniki mają istotny wpływ na zmienną zależną. Zmienne nieistotne (czyli w naszym przykładzie płeć) powinno się usunąć z modelu.

Przykład wprowadzający – wyniki końcowe... (3)

Wznawiamy analizę (*Ctrl + R*) i anulujemy aktualne okno, cofając się do miejsca, gdzie można wybrać zmienne do analizy. Tam na liście zmiennych niezależnych odznaczamy *Płeć*:

Mimo usunięcia jednej zmiennej z modelu współczynnik R^2 się niemal nie zmienił – model w 80,9% wyjaśnia zmienność końcowej sprawności pacjentów.

Błąd standardowy estymacji pozwala stwierdzić, iż model przybliża końcową sprawność pacjentów z dokładnością ±2,60 pkt

Podsumowanie regresji zmiennej zależnej: Poziom sprawności (po rehabilita
R= ,89956592 R^2= ,80921884 Popraw. R2= ,80844952
F(2,496)=1051,9 p<0,0000 Błąd std. estymacji: 2,6007

N=499	b*	Bł. std. z b*	b	Bł. std. z b	t(496)	p
W. wolny			6,0485	0,732378	8,25875	0,0000
Wiek	-0,06	0,020830	-0,0305	0,009923	-3,07324	0,0022
Poziom sprawności (przed rehabilitacją)	0,88	0,020830	0,8525	0,020272	42,05396	0,0000

W kolumnie „B" podane są wartości współczynników modelu, który, jak widać, przyjmuje ostatecznie postać:

Końcowy poziom sprawności = 6,0459 – 0,0305 · Wiek + + 0,8525 · Wyjściowy poziom sprawności

Możemy dokonać interpretacji parametrów:

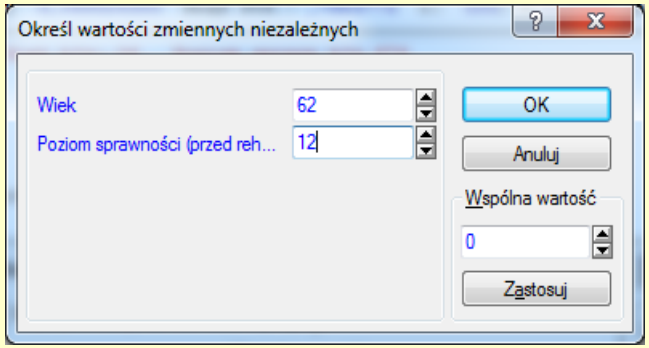
- **u pacjenta starszego o rok oczekiwany wynik końcowy jest niższy o 0,03 pkt (10 lat – to spadek rzędu 0,3 pkt);**
- **u pacjenta z wyższą o 1 pkt wyjściową sprawnością, końcowy wynik będzie – średnio rzecz biorąc – wyższy o 0,85 pkt.**

Wszystkie zmienne w modelu są istotne statystycznie.

Przykład wprowadzający – …i prognoza (4)

Wznawiamy analizę i w zakładce *Reszty, założenia, predykcja* możemy dokonać prognozy wyniki końcowego rehabilitacji dla nowoprzyjętego pacjenta – na przykład w wieku 62 lat i o wyjściowej sprawności 12 pkt.

Prognozę wyznaczymy wraz z tzw. przedziałem predykcji na poziomie ufności 80%. W tym celu w polu „Alfa" wprowadzamy wartość 0,2 (błąd prognozy) i zaznaczamy opcję Oblicz granice predykcji, po czym wywołujemy okno wprowadzania wartości zmiennych niezależnych



Zmienna	Obliczanie wartości (Efekty rehabilitacji zmiennej: Poziom sprawności (po rehal		
	Wagi b	Wartość	Wagi b *Wartość
Wiek	-0,030494	62,00000	-1,89064
Poziom sprawności (przed rehabilitacją)	0,852506	12,00000	10,23007
W. wolny			6,04852
Przewidyw.			**14,4**
-80,0%GP			**11,0**
+80,0%GP			**17,7**

W oknie wynikowym otrzymujemy szczegółowe wyliczenia wraz z wynikami końcowymi:

- ***prognozą punktową (wartość przewidywana): 14,4 pkt;***
- ***zakresem 80% przedziału predykcji : 11,0-17,7 pkt;***
- ***wynik można zapisać w taki sposób: 14,4 (11,0-17,7) pkt;***
- ***przedział predykcji to zakres, w którym z określonym prawdopodobieństwem (w tym przykładzie 80%) powinna znaleźć się wartość prognozowana.***

Najważniejsze wyniki analizy regresji

Współczynnik determinacji R^2 – określa procent zmienności cechy zależnej wyjaśnianej przez model. Tak więc jest to miernik jakości dopasowania modelu do danych i jako taki może służyć do porównywania kilku modeli i wyboru najlepszego. Współczynnik determinacji przyjmuje wartości od 0 do 100%, przy czym oczywiście im jego wartość jest większa tym model lepiej dopasowany.

UWAGA! Współczynnik R^2 rośnie wraz ze zwiększaniem liczby zmiennych w modelu. Gdybyśmy więc, jako jedyne kryterium jakości dopasowania, przyjęli jego wartość, wprowadzimy do modelu wszystkie dostępne w bazie danych cechy objaśniające. W ten sposób otrzymamy co prawda model najlepiej dopasowany, lecz jego złożoność nie pozwoli wyciągnąć sensownych wniosków praktycznych, ponadto wzajemne oddziaływania licznych zmiennych niezależnych zaburzają najczęściej ich relacje z cechą zależną.

Dlatego też należy wziąć pod uwagę istotność statystyczną zmiennych w modelu.

Istotność statystyczna zmiennych

Prawdopodobieństwo testowe *p* dla zmiennych występujących w modelu – każde dane da się „wyjaśnić” jeżeli do modelu regresji wprowadzi się bardzo dużo zmiennych niezależnych.

Jednak aby określić, czy poszczególne zmienne w modelu regresji opisują jakąś istotną część zmienności cechy zależnej (*Y*), przeprowadza się odpowiednie testy statystyczne.

W szczególności poddaje się weryfikacji zerową hipotezę, według której wkład danej zmiennej w wyjaśnianie zmienności cechy *Y* jest nieistotny.

Wynikiem testu statystycznego jest prawdopodobieństwo testowe *p*, którego niskie wartości pozwalają odrzucić „nieciekawą” hipotezę o braku znaczenia zmiennej objaśniającej w modelu (zwyczajowo za istotne statystycznie przyjmuje się wartości $p < 0,05$).

Prognozowanie na podstawie modelu regresji

Przewidywanie wartości zmiennej zależnej dla ”nowych” wartości zmiennych niezależnych ma sens, gdy model jest dobrze dopasowany, to znaczy wartość współczynnika determinacji jest wysoka (brak nietstey zgodności, co do kryterium „dobrego” dopasowania – najczęściej przyjmuje się, że R^2 powinien być większy niż 80%).

Jak zawsze w statystyce prognoza musi być obarczona pewnym błędem. Miarą jakości prognozy jest tzw. poziom ufności (standardowo przyjmowana jego wartość to 95% = 0,95).

Przedział dla oceny wartości przeciętnych zmiennej zależnej nazywany jest przedziałem ufności, a dla konkretnej jednostki statystycznej przedziałem predykcji. Przedział predykcji jest zawsze szerszy od przedziału ufności.

Na tych zajęciach przyjmiemy, że w modelach dotyczących szeregów czasowych będziemy wyznaczać przedziały ufności.

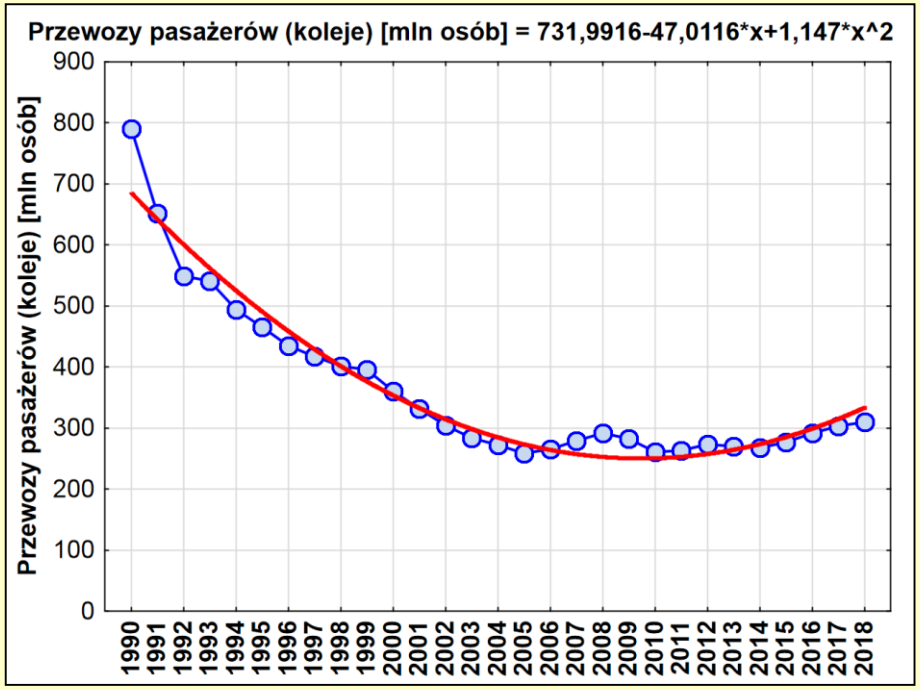
Modele regresji w prognozowaniu zjawisk czasowych

Analiza dotyczy danych z pliku *Transport w Polsce 1990-2018* o *liczbie pasażerów przewożonych koleją*. Celem analizy będzie sporządzenie prognozy tej wielkości dla alternatywnych modeli tej na lata 2019-2022.

Analizę *danych czasowych* rozpoczynamy *ZAWSZE* od prezentacji graficznej danego zjawiska.

Graficzna wizualizacja danych

Wykorzystując możliwość dopasowania pewnych modeli trendu bezpośrednio na wykresie liniowym, sporządzono graficzną prezentację dopasowania do danych rzeczywistych trendu kwadratowego.



Widocznym jest, że model liniowy w ogóle nie będzie pasował do analizowanych danych, dlatego prognozę wykonamy za pomocą modelu kwadratowego.
Lub za pomocą bardziej złożonej funkcji o podobnym kształcie

Model regresji – przygotowanie danych i wybór zmiennych do analizy

Aby przeprowadzić analizę regresji, w arkuszu danych musi pojawić się zmienna zawierająca informacje o numerze okresu czasowego (oraz jej ewentualne przekształcenia).

Jeżeli chcemy dopasować do danych funkcję kwadratową, to w arkuszu należy dodać dwie zmienne pomocnicze: *X* – *z numerami obserwacji (=v0)* oraz *X2* – *z kwadratami zmiennej X* ($=X^2$). Czynimy to za pomocą podanych w nawiasach *formuł*.

Następnie otwieramy okno analizy *Regresja wieloraka*, po czym jako *zmienną zależną* wskazujemy *Przewozy pasażerów koleją*, zaś jako *zmienne niezależne* OBIE ZMIENNE pomocnicze – *X* i X^2.

	16 X	17 X2
1990	1	1
1991	2	4
1992	3	9
1993	4	16
1994	5	25
1995	6	36
1996	7	49
1997	8	64
1998	9	81
1999	10	100
2000	11	121
2001	12	144
2002	13	169
2003	14	196
2004	15	225
2005	16	256
2006	17	289
2007	18	324
2008	19	361
2009	20	400
2010	21	441
2011	22	484
2012	23	529
2013	24	576
2014	25	625
2015	26	676
2016	27	729
2017	28	784
2018	29	841

Analiza regresji – podstawowe wyniki

Po przejściu do okna *WYNIKI REGRESJI WIELORAKIEJ* w zakładce *PODSTAWOWE* wywołujemy *PODSUMOWANIE: WYNIKI REGRESJI*.

Wartość współczynnika determinacji R^2, podawana jest zwyczajowo w procentach. Model w 95,6% opisuje zmienność liczby pasażerów w transporcie kolejowym w latach 1990-2018, a więc jest znakomicie dopasowany do danych

Błąd standardowy estymacji pozwala stwierdzić, iż rzeczywista liczba pasażerów kolei odstaje od wartości modelowanej o ±29,04 tys. osób

Podsumowanie regresji zmiennej zależnej: Przewozy pasażerów (koleje) [
R= ,97752322 R^2= ,95555165 Popraw. R2= ,95213255
F(2,26)=279,47 p<,00000 Błąd std. estymacji: 29,040

N=29	b*	Bł. std. z b*	b	Bł. std. z b	t(26)	p
W. wolny			731,99	17,36157	42,1616	0,0000
X	-3,02	0,171123	-47,01	2,66760	-17,6232	0,0000
X2	2,27	0,171123	1,15	0,08629	13,2932	0,0000

W kolumnie „B" podane są wartości współczynników modelu, który przyjął postać:

$$Y = 731{,}99 - 47{,}01 \cdot X + 1{,}15 \cdot X^2$$

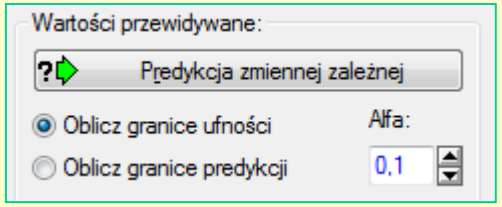
Jak łatwo sprawdzić, jest to oczywiście ten sam wzór, który otrzymaliśmy dopasowując model kwadratowy za pomocą wykresu liniowego i oczywiście PROGNOZY PUNKTOWE też będą identyczne.

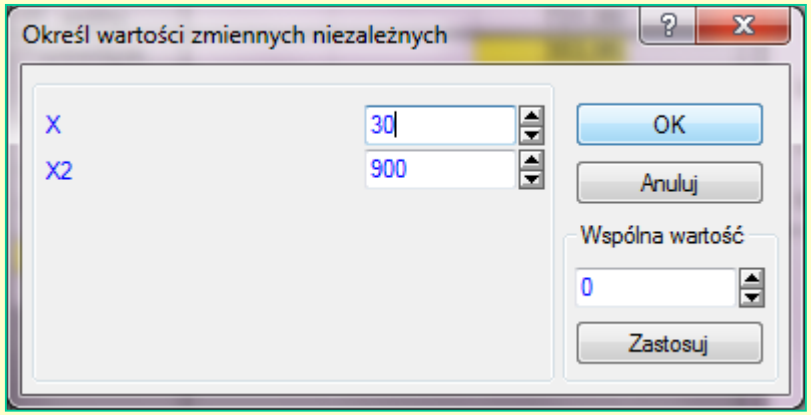
Wartości prawdopodobieństwa testowego p pozwalają na stwierdzenie, iż obie zmienne niezależne – X i X^2 są w statystycznie istotny sposób powiązane z liczbą pasażerów.

Analiza regresji – prognoza

W zakładce *RESZTY, ZAŁOŻENIA, PREDYKCJA* znajdują się narzędzia umożliwiające wyznaczenie punktowej i przedziałowej prognozy zmiennej *Y* dla zadanych wartości zmiennej *X* (w rozważanym przykładzie – liczby pasażerów przewożonych koleją dla kolejnych lat).

Aby wyznaczyć prognozę dla roku 2019 sprawdzamy w arkuszu danych jaki numer miałby ten rok w naszej bazie danych. Na tej podstawie wprowadzamy w pole *X* wartość *30*, zaś w pole *X2* – wartość *900*.





Zmienna	Obliczanie wartości (Transport w Pols zmiennej: Przewozy pasażerów (kolej		
	Wagi b	Wartość	Wagi b *Wartość
X	-47,0116	30,0000	-1410,35
X2	1,1470	900,0000	1032,30
W. wolny			731,99
Przewidyw.			353,95
-90,0%GU			324,33
+90,0%GU			383,56

Bardziej skomplikowany model (1)

Każdy model postaci:

$$Y_t = b_0 + b_1 \cdot f_1(t) + \ldots + b_k \cdot f_k(t) + e_t$$

jest, z punktu widzenia analizy regresji, tylko pozornie nieliniowy.

Ponieważ model kwadratowy zakłada dość szybkie tempo wzrostu, jako czynnik korygujący wprowadzimy doń funkcję hiperboliczną.

Model przyjmie postać:

$$Y_t = b_0 + b_1 X + b_2 X^2 + b_3 / X + e_t$$

Szacowanie parametrów tego modelu wygląda analogicznie jak funkcji kwadratowej, tylko wcześniej w arkuszu danych „dokładamy” jeszcze jedną pomocniczą zmienną – *1/X* (wyliczając ją za pomocą takiej właśnie formuły).

Bardziej skomplikowany model (2)

Po przejściu do okna *WYNIKI REGRESJI WIELORAKIEJ* w zakładce *PODSTAWOWE* wywołujemy *PODSUMOWANIE: WYNIKI REGRESJI*.

Model w 98,9% opisuje zmienność liczby pasażerów w transporcie kolejowym w latach 1990-2018, a więc jest znakomicie dopasowany do danych – jeszcze lepiej niż funkcja kwadratowa.

Błąd standardowy estymacji jest niemal dwa razy niższy niż dla funkcji kwadratowej ±14,64 tys. osób.

R= ,99455615 **R^2= ,98914194** Popraw. R2= ,98783898
F(3,25)=759,15 p<0,0000 Błąd std. estymacji **14,637**

N=29	b*	Bł. std. z b*	b	Bł. std. z b	t(25)	p
W. wolny			596,27	17,74179	33,6079	0,0000
X	-2,12	0,133493	-33,04	2,08100	-15,8788	0,0000
X2	1,57	0,117494	0,79	0,05924	13,3891	0,0000
1/X	0,33	0,037478	223,81	25,44948	8,7943	0,0000

Wartości prawdopodobieństwa testowego p pozwalają na stwierdzenie, iż WSZYSTKIE zmienne niezależne są istotne statystycznie.

W kolumnie „B" podane są wartości współczynników modelu, który przyjął postać:

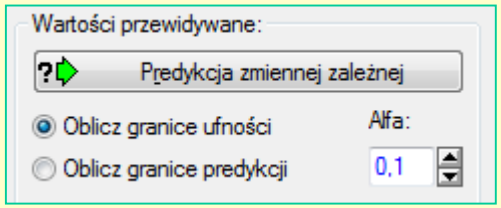
$$Y = 596{,}27 - 33{,}04 \cdot X + 0{,}79 \cdot X^2 + 223{,}81 \cdot 1/X$$

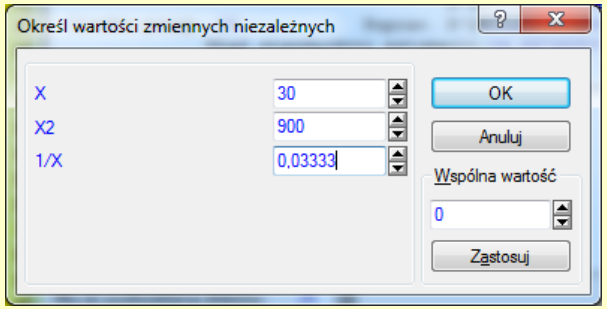
Takiego modelu nie da się już dopasować do danych, za pomocą trendów wbudowanych w narzędzia graficzne programu STATISTICA (ani Excel).

Bardziej skomplikowany model (3)

Aby wyznaczyć prognozę dla roku 2019 sprawdzamy w arkuszu danych jaki numer miałby ten rok w naszej bazie danych. Na tej podstawie wprowadzamy w pole *X* wartość *30*, zaś w pole X^2 – wartość *900*, zaś w pole *1/X* – wartość *0,03333* (w zaokrągleniu).

Przyjmiemy 90% poziom ufności, jak dla modelu kwadratowego.





Zmienna	Obliczanie wartości (Transport w Polsc zmiennej: Przewozy pasażerów (kolej		
	Wagi b	Wartość	Wagi b *Wartość
X	-33,0437	30,0000	-991,31
X2	0,7932	900,0000	713,90
1/X	223,8101	0,0333	7,46
W. wolny			596,27
Przewidyw.			326,31
-90,0%GU			310,43
+90,0%GU			342,19

Zestawienie prognoz

Poniżej zestawiono prognozy liczby pasażerów w transporcie kolejowym na lata 2019-2022, uzyskane za pomocą modelu kwadratowego i kwadratowego ze składnikiem 1/*X*.

W tabeli podano wartości prognoz wraz z 90% przedziałem ufności.

Rok	Liczba pasażerów (w mln)	
	Model kwadratowy	Model kwadratowy ze składnikiem 1/X
2019	353,9 (324,3-383,6)	326,3 (310,4-342,2)
2020	376,9 (343,1-410,7)	341,4 (323,0-359,8)
2021	402,2 (363,8-440,5)	358,1 (336,9-379,3)
2022	429,7 (386,4-473,0)	376,4 (352,2-400,6)

Jak widać, wprowadzenie do modelu komponentu 1/*X*, zgodnie z oczekiwaniami, „spłaszczyło” prognozę. Ponadto w modelu tym, jako lepiej dopasowanym do danych historycznych, mamy zdecydowanie węższe przedziały ufności, czyli prognoza jest bardziej precyzyjna.

Uwagi końcowe – wiarygodność modelu i prognoz

Analizując otrzymane wyniki, należy pamiętać, iż zostały one uzyskane jedynie na podstawie informacji zawartych w wyjściowym szeregu czasowym – nie uwzględniono żadnych czynników zewnętrznych.

Rozważając wiarygodność prognoz należałoby uwzględnić:

- **perspektywy rozwoju ruchu turystycznego;**
- **ceny paliw i samochodów oraz rozwój infrastruktury drogowej – czyli rozwój konkurencyjnego środka transportu;**
- **zmiany demograficzne – spadek liczności populacji i jej starzenie się;**
- **inwestycje w tabor i infrastrukturę kolejową – mogą mieć zarówno pozytywny, ale też czasowo negatywny (wyłączenie pewnych tras), wpływ na liczbę pasażerów.**

Analiza reszt – informacja

Model regresji powinien być poddany jeszcze dokładniejszej analizie, przed wykorzystaniem go jako narzędzia do sporządzania prognoz.

Na jednym z kolejnych wykładów, poświęconych metodom ESTYMACJI NIELINIOWEJ – przedstawione zostaną podstawowe elementy tak zwanej analizy reszt.