

Prognozowanie i symulacje (lab. 5)

Analiza regresji w prognozowaniu

Wstęp

Do tej pory posługiwaliśmy się tylko i wyłącznie informacjami otrzymanymi za pomocą wykresów liniowych lub wykresów rozrzutu z dopasowaniem odpowiednich modeli. Bardziej uniwersalne narzędzie do konstruowania modeli trendów (z większą liczbą wyników i szerszymi możliwościami jeśli chodzi o zakres modeli) oferuje moduł **Regresja wieloraka**.

Za pomocą tego narzędzia można konstruować modele postaci:

$$Y = b_0 + b_1 \cdot X_1 + \dots + b_k \cdot X_k + e$$

(Y jest zmienną zależną, X_1, \dots, X_k – to zmienne niezależne, a e – oznacza błąd modelu, bowiem mamy do czynienia nie z zależnościami deterministycznymi, ale statystycznymi)

Poprzez konstruowanie dodatkowych zmiennych w arkuszu danych, bazując na modelu liniowym, można oszacować równania modeli w postaci pozornie nieliniowej, np. takich jak:

$$Y = a + b \cdot X + c \cdot X^2 + d \cdot X^3 \text{ (i ogólnie dowolnych wielomianów)}$$

$$Y = a + b/X \text{ (model hiperboliczny – jako zmienną dodatkową należy wstawić } 1/X)$$

$$\text{i inne ogólnej postaci: } Y = b_0 + b_1 \cdot f_1(X) + \dots + b_k \cdot f_k(X) + e$$

Tak więc, w najprostszej ujęciu, zmiennymi niezależnymi w naszych modelach będzie zmienna czasowa (numer obserwacji) i/lub jej proste przekształcenia.

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

UWAGA! Ponieważ dane w latach 2020, 2021 i 2022 były nieporównywalne (lockdown!) proszę zapisać plik pod nową nazwą, usunąć trzy ostatnie przypadki i wykonywać prognozy z perspektywy roku 2019. Dotyczy to przykładu 1, 2 i 3.

Celem analizy będzie skonstruowanie prognozy przewozów koleją na kolejne lata za pomocą modelu trendu liniowego. W arkuszu danych (najlepiej na końcu) wstawiamy dodatkową zmienną X i wypełniamy ją numerami obserwacji ($=v0$).

Za pomocą polecenia **Statystyka / Regresja wieloraka** uruchamiamy wejściowe okno służące do konstruowania modeli liniowych i wybieramy zmienne: na liście **zmiennych zależnych** wskazujemy **Przewozy pasażerskie koleją** na liście zmiennych niezależnych wprowadzając uprzednio pomocniczą zmienną X (zawierającą numer obserwacji). Po zatwierdzeniu wyboru zmiennych przechodzimy do okna **Wyniki regresji wielokrotnej**, gdzie w zakładce **Podstawowe** za pomocą przycisku **Podsumowanie: wyniki regresji** wywołujemy najważniejsze wyniki analiz. Na razie w arkuszu wyników interesować nas będą dwie kolumny **b** i **p** oraz wartość **R²** w nagłówku tabeli.

Podsumowanie regresji zmiennej zależnej: Przewozy pasażerów (kol)						
R= ,7806037 R^2= ,60934228 Popraw. R2= ,59539021						
F(1,28)=43,674 p<,000000 Błąd std. estymacji: 83,035						
N=30	b*	Bł. std. z b*	b	Bł. std. z b	t(28)	p
W. wolny			543,6	31,09429	17,48229	0,0000
X	-0,780604	0,118119	-11,6	1,75150	-6,60863	0,0000

W kolumnie **b** umieszczone są współczynniki modelu liniowego zaś w kolumnie **p** znajduje się ocena istotności poszczególnych składników modelu – jeżeli są one poniżej 0,05 oznacza to, że obecność danego czynnika w modelu jest uzasadniona. W nagłówku arkusza wyników znajduje się wartość **R²**, zwana **współczynnikiem determinacji**, którą na ogół wyraża się w procentach (65,3%). W naszym przypadku można stwierdzić, iż model liniowy jest przeciętnie dopasowany do danych. Współczynnik **R²** jest w nieco powyżej 60% wyjaśnia zmienność cechy zależnej.

Aby skonstruować prognozę dla kolejnych lat, wznawiamy analizę (Ctrl + R) i przechodzimy do zakładki **Reszty, założenia, predykcja**. Ponieważ teraz będziemy wyznaczać nie tylko prognozę punktową, ale także otaczający ją przedział ufności, musimy ustalić poziom zaufania do prognozy. W tym celu w odpowiednie pole (rysunek obok) wpisujemy **poziom błędu prognozy** – przykładowo, jeżeli chcemy otrzymać zakres 90% przedziału ufności dla prognozy wtedy wpisujemy poziom błędu 0,10).

Predykcja zmiennej zależnej

Oblicz granice ufności Alfa:

Oblicz granice predykcji

Po ustaleniu poziomu ufności klikamy przycisk **Predykcja zmiennej zależnej**. Podajemy odpowiedni numer prognozowanego okresu (dla roku 2020 – nr 31) i wywołujemy prognozę. W tabeli podana jest wartość prognozy punktowej i zakres okalającego ją przedziału ufności. Analogicznie sporządzamy prognozy dla roku 2021 i 2022. Podkreślimy, że wyniki obejmują nie tylko **prognozę punktową** (oczywiście identyczną z wynikami uzyskanymi wcześniej innymi metodami) ale także pewien przedział, w którym z **90% pewnością** powinna znaleźć się prognozowana wielkość. W rozważanym przykładzie szerokość przedziału prognozy jest dość duża, co obniża jej wiarygodność i praktyczne znaczenie. I każde podejść bardzo ostrożnie do otrzymanych wyników.

Obliczanie wartości (Transport w Pols zmiennej: Przewozy pasażerów (kolej)			
Zmienna	Wagi b	Wartość	Wagi b *Wartość
X	-11,5750	31,00000	-358,826
W. wolny			543,599
Przewidyw.			184,8
-90,0%GU			131,9
+90,0%GU			237,7

Rok	Przewozy pasażerów w transporcie kolejowych (w mln osób)	
	Model liniowy (R ² = 60,9%)	
	Prognoza punktowa	Prognoza przedziałowa (90% przedział ufności)
2020	184,8	131,9-237,7
2021		
2022		

Prognozowanie i symulacje (lab. 5)

Analiza regresji w prognozowaniu

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

Celem analizy będzie skonstruowanie prognozy **liczby autobusów** w Polsce na lata 2020-2023 za pomocą modelu **liniowego** i **kwadratowego**. Ponieważ wykorzystywać będziemy model kwadratowy w arkuszu danych dodajemy od razu dwie zmienne: X – zawierającą numery przypadków oraz X^2 – zawierającą kwadraty numerów obserwacji.

Konstruując prognozę za pomocą modelu liniowego jako zmienną niezależną wprowadzamy tylko zmienną X zaś w przypadku modelu kwadratowego **zarówno** X jak i X^2 . Prognozę przedziałową proszę sporządzić przy **80% poziomie ufności**.

UWAGA! Podstawiając wartości X i X^2 dla prognozy kwadratowej należy wstawić numer obserwacji (X) i kwadrat numeru obserwacji (X^2). Na przykład, prognozując liczbę autobusów dla roku 2020 należy przyjąć $X = 31$, zaś $X^2 = 961$.

Rok	Przewozy liczby zarejestrowanych autobusów (w tys.)			
	Model liniowy ($R^2 =$ %)		Model kwadratowy ($R^2 =$ %)	
	Prognoza punktowa	Prognoza przedziałowa (80% przedział ufności)	Prognoza punktowa	Prognoza przedziałowa (80% przedział ufności)
2020				
2021				
2022				
2023				

Przykład 3 (Transport w Polsce 1990-2022 (R))

Podjmiemy próbę skonstruowania prognozy **liczby pasażerów przewożonych koleją** za pomocą modelu postaci:

$$Y = a + b \cdot X + c \cdot X^2 + d/X$$

W arkuszu danych dodajemy odpowiednie zmienne pomocnicze (X , X^2 oraz $1/X$) i wyliczamy ich wartości za pomocą odpowiednich formuł. Następnie wyznaczamy współczynniki modelu (analogicznie jak w poprzednich przykładach) i oceniamy ich istotność statystyczną (wartości prawdopodobieństwa testowego p powinny być poniżej 0,05). Jeżeli współczynniki są istotne, wyznaczamy współczynnik determinacji (R^2) oraz prognozę punktową i przedziałową (na 85% poziomie ufności) na lata 2020-2022.

Proszę podać znaleziony wzór funkcji, za pomocą której wyznaczano prognozy:

$Y = \dots\dots\dots$

Rok	Przewozy pasażerów w transporcie kolejowym (w mln osób) za pomocą funkcji postaci: $a + b \cdot X + c \cdot X^2 + d/X$	
	$R^2 =$ %	
	Prognoza punktowa	Prognoza przedziałowa (85% przedział ufności)
2020		
2021		
2022		

W oryginalnym zbiorze danych są informacje o przewozach pasażerów koleją w 2020, 2021 i 2022 r. Proszę porównać je z wyznaczoną prognozą (Y_p), wyznaczając błąd procentowy prognozy za pomocą wzoru:

Błąd procentowy = $(Y - Y_p)/Y \cdot 100\%$

1) Co oznacza ta informacja – czy prognoza była za niska czy za wysoka?

2) Jak bardzo spadły przewozy kolejowe wskutek „pandemii” i związanych z nią ograniczeń normalnego życia społecznego w stosunku do wartości prognozowanej przy założeniu kontynuacji trendu wzrostowego.

Powstaje pytanie, jak skonstruować wykres pokazujący oryginalne dane i przebieg dopasowanej funkcji. W analizie regresji nie ma wbudowanego narzędzia tworzenia wykresów (dlatego, że narzędzie to używane jest nie tylko do danych czasowych), ale można bez problemu wyznaczyć wartości modelowanej funkcji i wkleić je do oryginalnego arkusza danych. W tym celu proszę wznowić analizę i w zakładce *Reszty, założenia, predykcja* wybrać polecenie *Wykonaj analizę reszt*, a następnie za pomocą przycisku *Podsumowanie* wywołać tabelę z wartościami oryginalnymi, dopasowanymi, resztami i innymi statystykami.

Proszę skopiować wartości z kolumny nr 2 (*Przewidywane wartości*) i wkleić je do nowej kolumny w oryginalnym arkuszu danych. Proszę tę nową kolumnę nazwać *Model prognostyczny*.

Za pomocą *Wykresu liniowego* w wersji *Wielokrotnej* proszę skonstruować wykres pokazujący oryginalne dane dotyczące liczby przewozów pasażerów koleją oraz dopasowany do nich model prognostyczny.

Wykres proszę sformatować według reguł poznanych na wcześniejszych zajęciach.

UWAGA: Gdyby do wyjściowego arkusza dodać trzy nowe przypadki i wprowadzić w kolumnie *Model prognostyczny* wyznaczone powyżej prognozy, na wykresie poza modelem pojawiłyby się też wartości prognozowane na kolejne lata.