AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ ZARZĄDZANIA

Projekt zaliczeniowy

Metody Ekonometryczne

Budowa oraz diagnostyka modelu ekonometrycznego na podstawie zbioru danych nr. 5

Autor: *Justyna Krok*, *Artur Karamon*
Kierunek studiów: *Informatyka i Ekonometria*
Prowadzący: *mgr. Aneta Bech*

Kraków, 2019

Spis treści

1. Wprowadzenie

Celem projektu jest analiza danych ze zbioru 5 z wykorzystaniem znanych metod ekonometrycznych. Projekt w głównej mierze opiera się na stworzeniu modelu regresji liniowej oraz wyestymowaniu parametrów modelu metodą najmniejszych kwadratów.

Na wstępie badania zostały przedstawione dane oraz szczegółowa analiza podstawowych statystyk. Dzięki temu możliwe jest lepsze zrozumienie danych oraz zależności w nim występujących. Ważną częścią jest wizualizacja zależności oraz rozkładu cech statystycznych.

W kolejnej części projektu dokonano transformacji na zbiorze danych, które są konieczne do dalszej analizy. Za pomocą znanych metod dobrano zmienne do modelu. Konieczne było zbadanie podstawowych założeń metody najmniejszych kwadratów oraz diagnostyka końcowego modelu.

W końcowej części pracy wykorzystano stworzony model w celu prognozowania wartości zmiennej objaśnianej dla próbki testowej. W celu weryfikacji poprawności predykcji wykorzystano miary błędów predykcji ex psot.

Regresja liniowa w głównej mierze opiera się na znalezieniu liniowego dopasowania modelu do danych. Warto pamiętać, że zależności w danych nie zawsze są liniowe i metody regresji liniowej mogą nie być odpowiednie dla badanego zbioru danych. Istnieje kilka metod szacowania parametrów modelu regresji liniowej jak np. metoda największej wiarygodności czy metoda najmniejszych kwadratów.

Metoda najmniejszych kwadratów jest najczęściej stosowana w regresji liniowej. Oszacowane parametry pozwalają na zbadanie zależności między zmienną objaśnianą, a zmiennymi objaśniającymi. Oprócz kierunku zależności MNK pozwala oszacować również dokładny wpływ danej cechy na wartość zmiennej objaśnianej. Poprawny model pozwala również na predykcję, a więc na przewidywanie wartości zmiennej objaśnianej na podstawie znajomości zmiennych objaśnianych.

2. Badanie empiryczne

W tej części projektu przeprowadzono badanie empiryczne na zgromadzonych danych. Dokonano szczegółowej prezentacji i opisu danych. Obliczono podstawowe statystyki i zwizualizowano zależności między zmiennymi.

Dalsza część pracy zawiera podział danych na zbiór uczący oraz testujący. Zbiór uczący wykorzystany został do stworzenia modelu, a zbiór testujący do weryfikacji czy stworzony model może zostać wykorzystany do predykcji.

Rozdział ten zawiera opis wszystkich przeprowadzonych transformacji, opis stworzonego modelu oraz jego szczegółową diagnostykę.

2.1. Prezentacja i opis danych

Zgromadzone dane składają się z ośmiu zmiennych – jednej zmiennej objaśnianej oraz siedmiu zmiennych objaśniających. Zmienna X7 jest zmienną kategoryczną i przyjmuje wartości „A”, „B” oraz „C”. Wszystkie pozostałe zmienne są zmiennymi ilościowymi i przyjmują wartości rzeczywiste. Dane posiadają 1000 obserwacji. Poniżej znajduje się tabela przedstawiająca wartości zmiennych dla kilku pierwszych operacji.

X1	X2	X3	X4	X5	X6	X7	Y
-301,00	275,00	-353,00	-543,00	83,00	-1016,00	A	-581,00
-498,00	145,00	664,00	-351,00	-503,00	-1082,00	A	-596,00
-557,00	96,50	-136,00	-299,00	-163,00	-549,00	A	-521,00
-339,00	344,00	26,08	-182,00	364,00	-1012,00	B	-451,00
-633,00	645,00	-221,00	162,00	-28,10	-2606,00	B	-263,00
131,00	218,00	-93,80	240,00	87,30	-785,00	C	63,10
-305,00	371,00	659,00	-504,00	-245,00	-1730,00	B	-436,00
-398,00	304,00	102,00	-119,00	216,00	-999,00	B	-388,00
-153,00	132,00	111,00	-447,00	48,00	-481,00	A	-545,00
289,00	-388,00	343,00	76,90	-422,00	1130,00	C	97,50
-527,00	514,00	829,00	-234,00	103,00	-1953,00	B	-418,00
-372,00	502,00	-268,00	-687,00	-478,00	-2487,00	A	-725,00
…	…	…	…	…	…	…	…

W celu lepszego zrozumienia danych w poniższej tabeli przedstawiono podstawowe statystyki analizowanych zmiennych takie jak średnia, odchylenie standardowe, mediana skośność, kurtoza, współczynnik zmienności oraz liczbę braku danych w zbiorze.

zmienna	średnia	odchylenie standardowe	mediana	skośność	kurtoza	wsp. zmienności	liczba wartości NA
Y	-346.23	251.94	-346.38	0.04	-0.32	0.73	3
X1	-216.37	251.94	-206.31	-0.12	0.09	1.16	3
X2	175.06	262.83	179.42	-0.04	0.09	1.50	3
X3	46.48	259.04	42.50	-0.09	0.17	5.57	3
X4	-251.37	251.90	-251.29	-0.01	0.17	1.00	3
X5	-49.54	261.24	-53.00	0.06	0.16	5.27	3
X6	-748.70	1089.55	-774.07	0.05	0.00	1.46	3

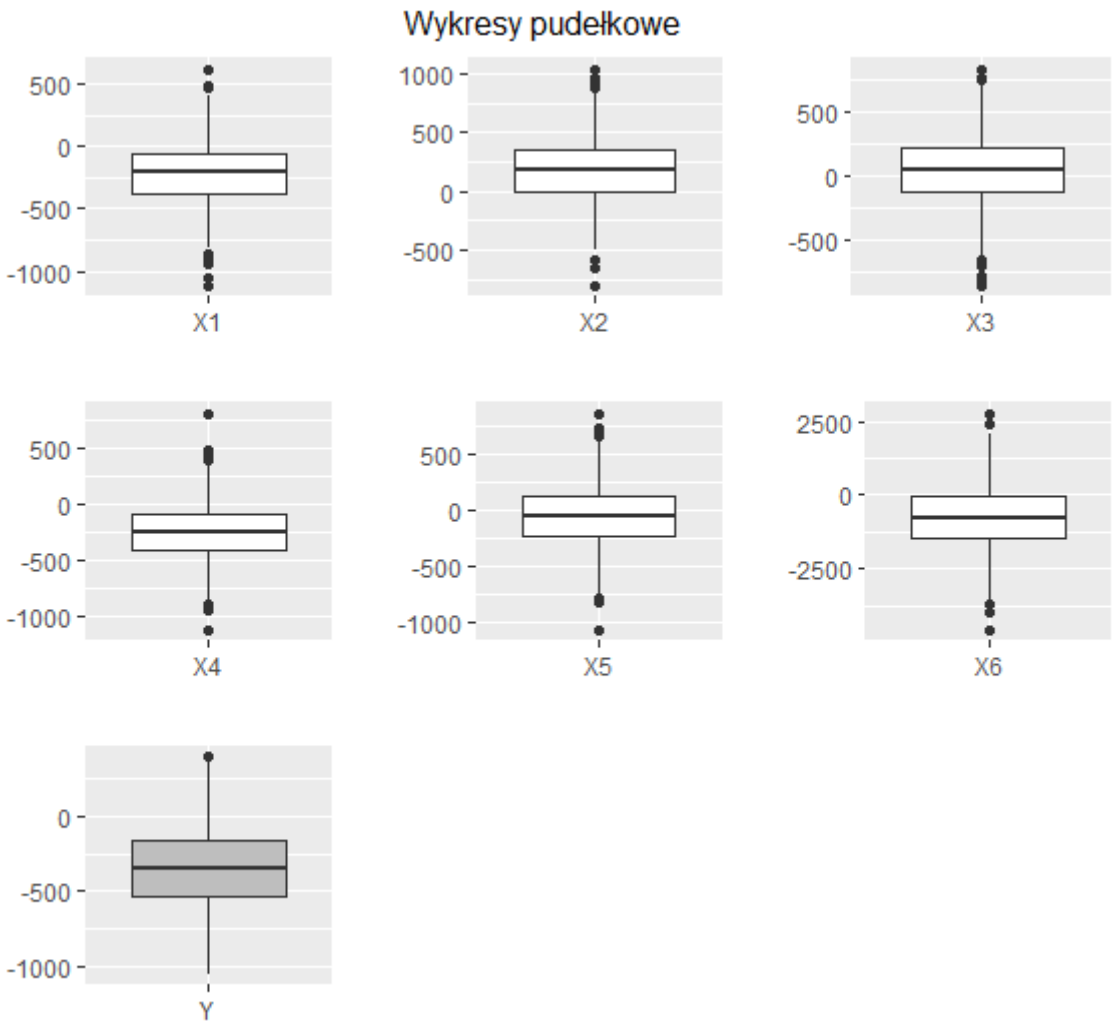
Każda ze zmiennych podsiada trzy braki w wartościach danych. Łącznie jest ich 24. W dalszej analizie konieczne jest ich usunięcia lub zastąpienie innymi wartościami gdyż mogą one negatywnie wpływać na otrzymane wyniki.

Zmienna Y jest zmienną objaśnianą, której wartości są głównie ujemne. Średnia zmiennej Y wynosi -346, a jej odchylenie 252. Odchylenie standardowe ma dużą wartość co świadczy o dużym rozproszeniu danych. Współczynnik zmienności jest miarą zróżnicowania rozkładu cechy i dla zmiennej Y wynosi -0.73 co potwierdza duże rozproszenie danych. Kurtoza jest miarą koncentracji, a jej ujemna wartość świadczy o małej koncentracji wyników wokół średniej. Mediana zmiennej Y jest bliska średniej co świadczy o małej asymetrii rozkładu, wartość współczynnika skośności potwierdza niską skośność prawostronną.

Zmienne X1, X2, X3, X4 oraz X5 mają bardzo podobne statystyki. Ich odchylenie standardowe jest wysokie i wynosi około 250-260. Należy zwrócić uwagę na współczynnik zmienności, który w porównaniu do odchylenia standardowego jest względną miarą zróżnicowania cechy i zależy od średniej. Zmienne X3 oraz X5 są zmiennymi najbardziej rozproszonymi, a ich współczynnik zmienności jest większy niż pięć. Współczynnik pozostałych zmiennych jest niewiele większy od 1. Rozproszenie zmiennych objaśnianych jest znacznie większe niż rozproszenie zmiennej objaśnianej. Zmienna X6 ma bardzo duże odchylenie standardowe jednak współczynnik zmienności pokazuje, że nie jest ona bardziej rozproszona niż pozostałe zmienne. Kurtoza zmiennych objaśniających jest nieujemna – wyniki są skoncentrowane blisko średniej jednak nie w znacznym stopniu (wartości kurtzy są niskie). Zmienna X6 posiada zerową kurtozę, a więc identyczną jak w rozkładzie normalnym.

Mediana każdej ze zmiennych objaśnianych jest bliska wartości średniej. Niskie współczynniki skośności wskazują na symetryczność rozkładów. Zmienna X1 oraz X4 osiadają nieznaczną lewostronną skośność, a pozostałe zmienne posiadają nieznaczną prawostronną skośność.

Zmienna kategoryczna X7 może przyjmować trzy różne wartości. Wszystkie rozpatrywane wartości występują w takiej samej liczbie w rozpatrywanych zbiorze danych.

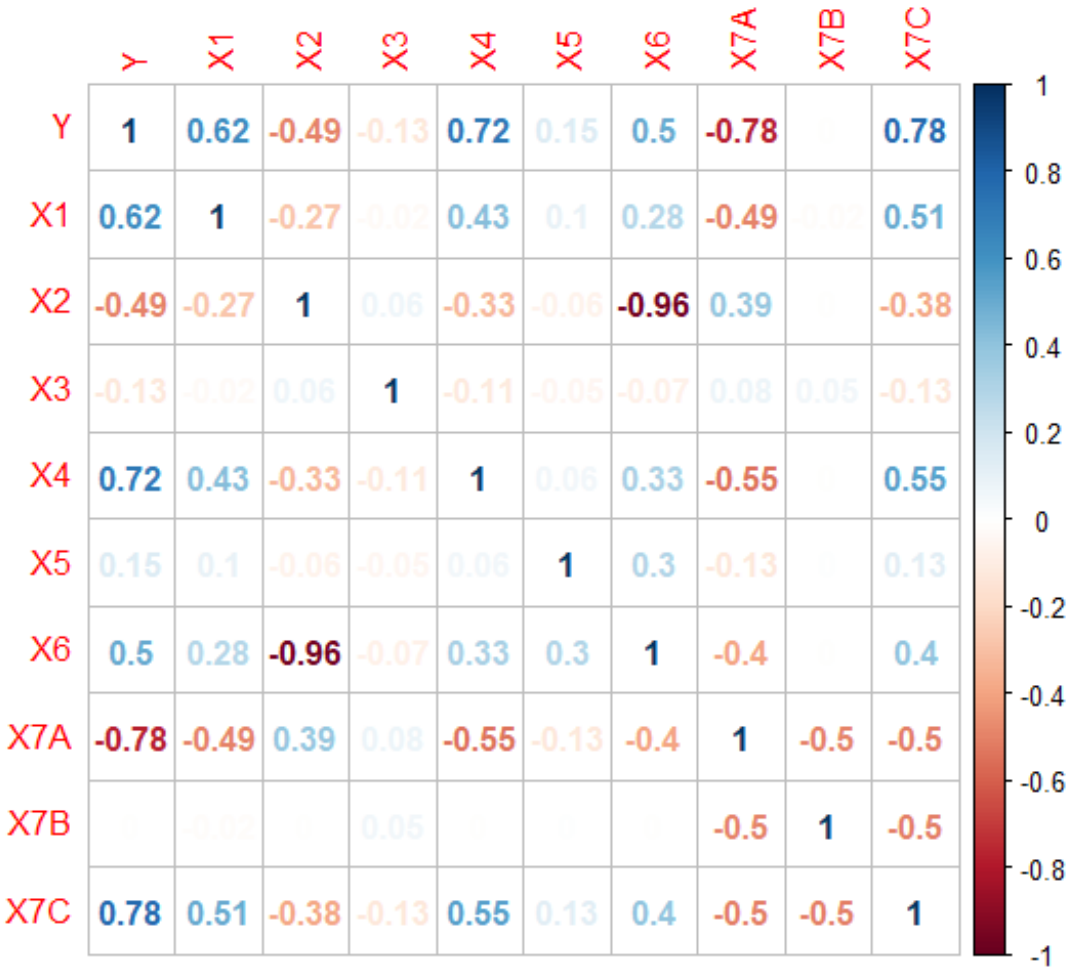


Wykres pudełkowy jest graficznym przedstawieniem rozkładu cechy statystycznej, wspomagającym proces analizy danych statystycznych. Wszystkie zmienne mają podobne rozkłady empiryczne. Analizując położenie wykresu możemy określić zakres danych oraz wskazać wartości minimalne oraz maksymalne. Największy zakres ma zmienna X6, która jednocześnie przyjmuje najmniejsze oraz największe wartości. Pozostałe wykresy mają podobne położenie.

Analizują wykresy można zauważyć, że większość zmiennych ma podobne rozproszenie, wyjątkiem jest zmienna X6, której rozproszenie jest kilkakrotnie razy większe, oraz zmienna Y, która ma węższy zakres niż pozostałe zmienne.

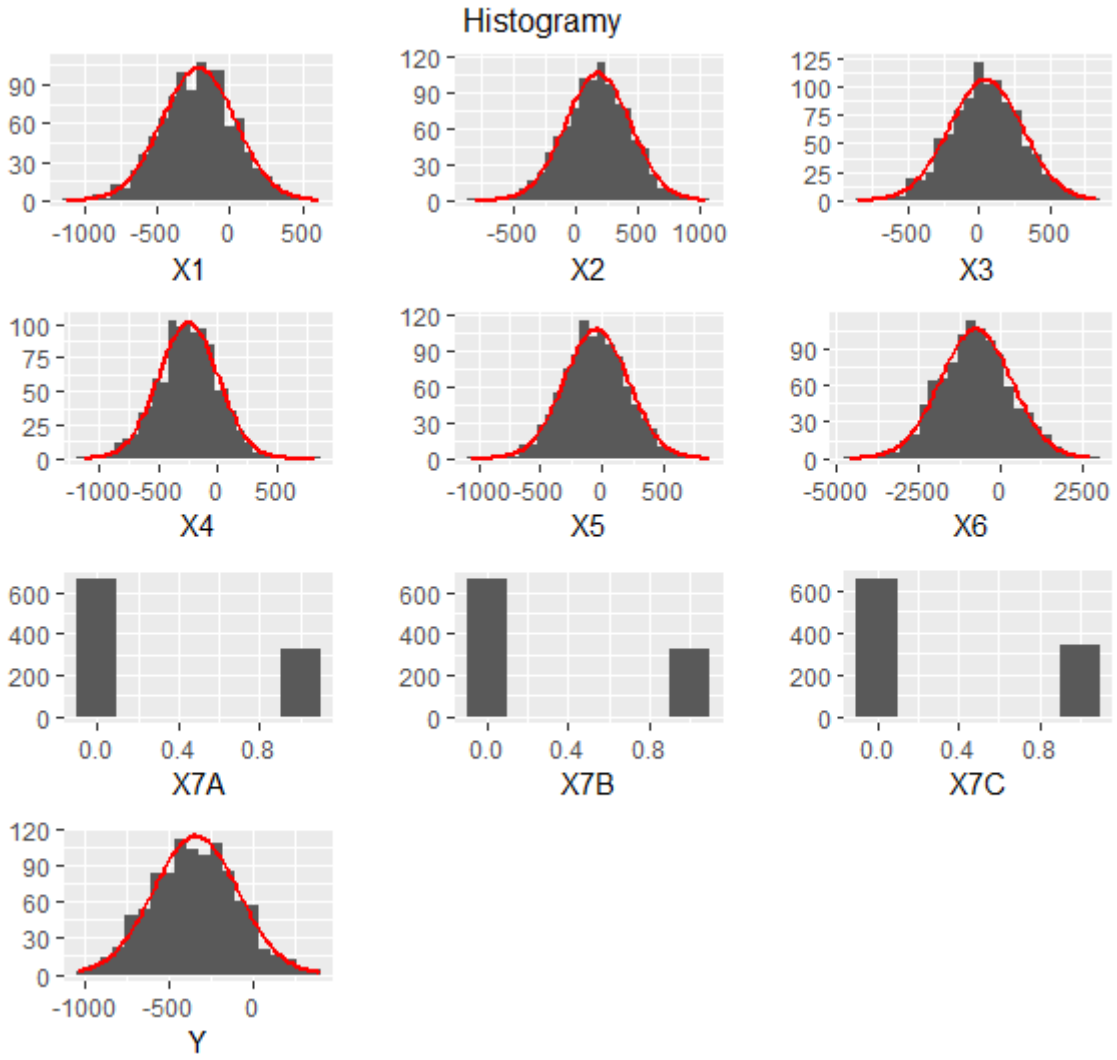
Mediana zmiennych leży na środku każdego pudełka, co wskazuje na symetryczność zmiennych. Jedynie w zmiennej X1 można zauważyć, że mediana znajduje się bliżej wartości maksymalnej co wskazuje na lewostronną asymetrię rozkładu. Nie jest to jednak wysoka wartość.

Zmienne objaśniające posiadają wartości odstające, które mogą zaburzać estymację modelu. Zmienna Y ma stosunkowo mało wartości odstających.



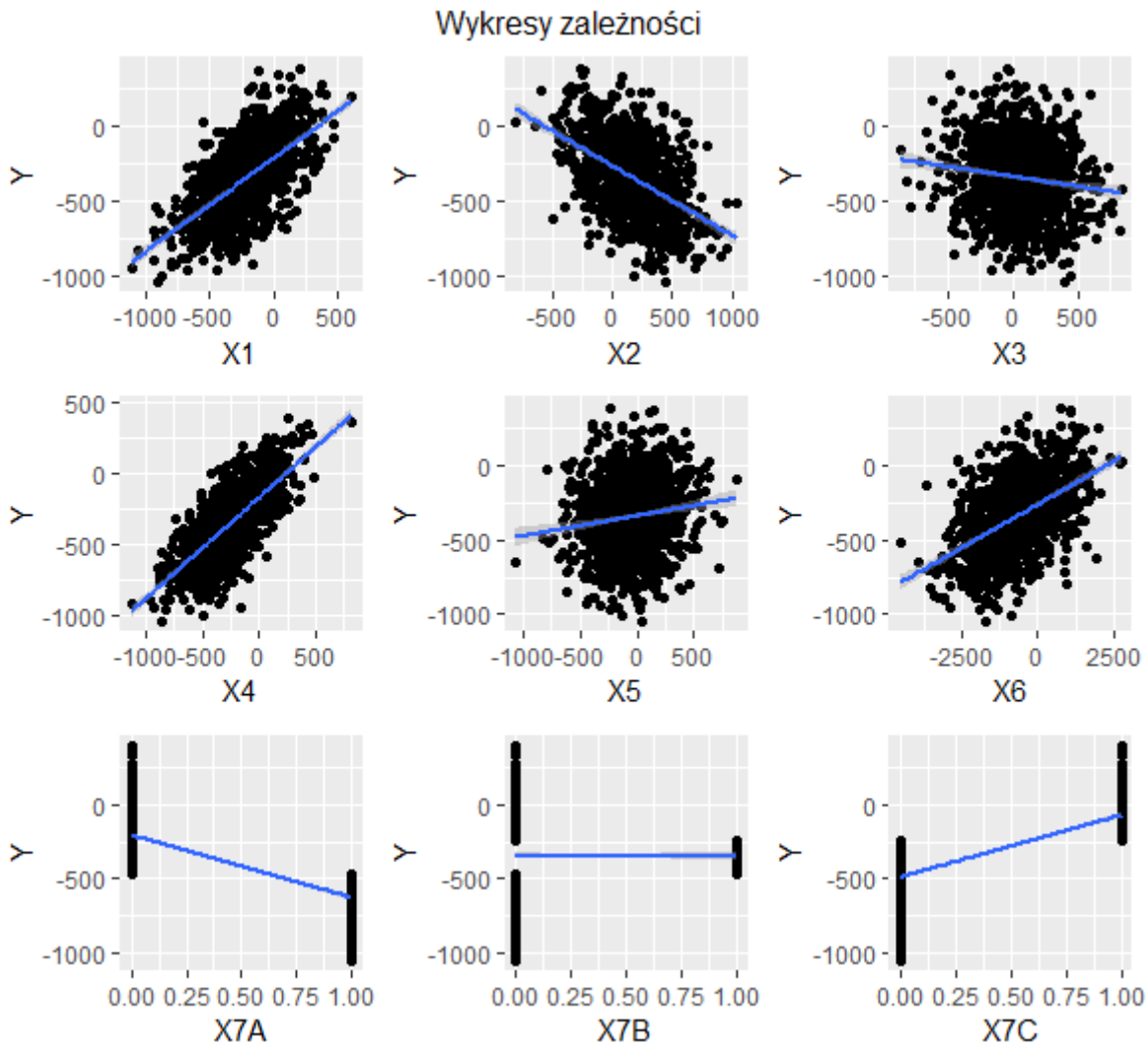
Powyżej znajduje się macierz korelacji analizowanych zmiennych. Zmienna Y jest wysoko skorelowana ze zmiennymi: X1, X4 oraz X7, średnio skorelowana ze zmiennymi: X2 oraz X6 oraz nisko skorelowana ze zmiennymi X3 oraz X5. Macierz korelacji pozwala na wstępny dobór zmiennych do modelu.

Biorąc pod uwagę korelacje między zmiennymi objaśniającymi można zauważyć wysoką wartość korelacji między zmienną X2 oraz X6, która wynosi aż -0,96. Oznacza to, że zmienne te przenoszą niemal tą samą informację, a wykorzystanie obu zmiennych w modelu może przyczynić się do błędnego oszacowania parametru. Pozostałe wartości korelacji nie są wysokie. Dokładne badanie współliniowości zmiennych prowadzono w dalszej części projektu.



Histogram jest wykresem, który przedstawia rozkład empiryczny cechy. Składa się z przedziałów oraz liczby obserwacji, które się w nich znajdują. Na przedstawionych wykresach zaznaczono na czerwono dystrybuantę rozkładu normalnego.

Rozkłady zmiennych ilościowych w znacznym stopniu przypominają rozkład normalny. Niskie wartości skośności oraz kurtozy potwierdzają przypuszczenie, że zmienne mają rozkłady normalne jednak warto zauważyć, że rozkłady te mają nieznacznie różne parametry – zwłaszcza zmienna X6 oraz Y.



W celu zweryfikowania zależności między zmiennymi objaśniającymi, a zmienną objaśnianą stworzono wykresy przedstawione. Wizualne przedstawienie zależności ułatwia interpretację.

Zmienna X4 w najlepszym stopniu obrazuje zależność liniową. Zależności zmiennych X2 oraz X6 również można określić jako liniowe, jednak funkcje, które je opisują mają przeciwne znaki przy współczynniku kierunkowym – potwierdza to ujemna korelacja między tymi zmiennymi.

Wykresy zależności zmiennych X3 oraz X5 przedstawiają skupisko danych, których opisanie funkcją liniową wydaje się mało dokładne. Ich współczynniki kierunkowe mają niskie wartości. Warto pamiętać, że zmienne te są nisko skorelowane ze zmienną objaśnianą Y.

Trzy ostatnie wykresy przedstawiają zależności zmiennej kategorycznej w trzech przypadkach (gdy przyjmuje wartość „A”, „B” oraz „C”) ze zmienną Y. Gdy zmienna X6 przyjmuje wartość A, widać, że zależność jest liniowa i malejąca, dla wartości C funkcja jest liniowa oraz malejąca, a dla wartości B – funkcja jest stała (brak zależności ze zmienną Y).

2.2. Podział próbki na zbiór uczący i testowy

Zgromadzone obserwacje podzielono na dwa zbiory danych: zbiór uczący i zbiór testowy w proporcjach 750:250. Podział został dokonany w sposób losowy. Zbiór uczący służy do stworzenia modelu oraz przeprowadzenia jego diagnozy. Z wykorzystaniem modelu prognozujemy wartości Y dla zbioru testowego, a następnie liczymy błędy oszacować porównując wartości rzeczywiste z wartościami prognozy.

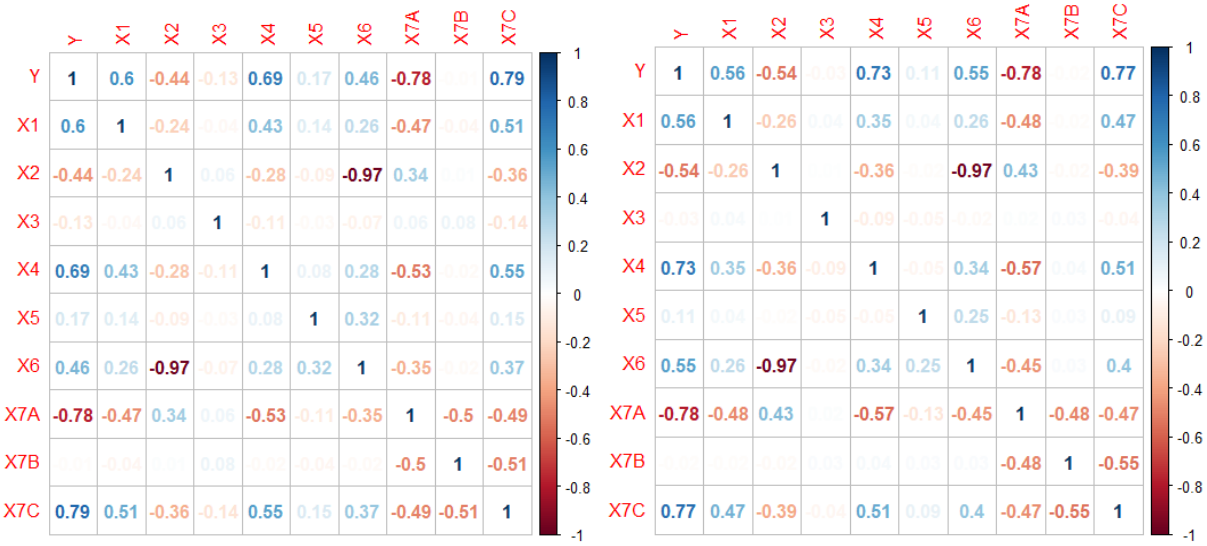
W celu identyfikacji, czy podział na dwa zbiory jest poprawny, obliczono podstawowe statystyki dla obu zbiorów oraz wygenerowano wykresy korelacji pomiędzy zmiennymi. Statystyki nie powinny mieć znaczących różnic.

zmienna	średnia	odchylenie standardowe	Mediana	skośność	kurtoza	wsp. zmienności	liczba wartości NA
Y	-347.18	243.66	-353.47	0.04	-0.44	-0.70	2
X1	-213.46	234.48	-206.72	0.01	-0.36	-1.10	1
X2	171.78	248.32	174.53	0.00	-0.33	1.45	1
X3	62.76	248.26	62.80	-0.05	-0.15	3.96	2
X4	-252.33	243.12	-248.47	-0.06	-0.30	-0.96	3
X5	-53.30	248.30	-53.23	0.05	-0.20	-4.66	2
X6	-743.17	1043.67	-755.24	0.02	-0.27	-1.40	3
X7A	0.33	0.47	0.00	0.73	-1.47	1.43	3
X7B	0.33	0.47	0.00	0.70	-1.51	1.41	3
X7C	0.34	0.47	0.00	0.68	-1.54	1.40	3

zmienna	średnia	odchylenie standardowe	mediana	skośność	kurtoza	wsp. zmienności	liczba wartości NA
Y	-330.52	231.02	-298.58	0.03	-0.42	-0.70	1
X1	-200.27	246.23	-191.29	-0.02	-0.52	-1.23	2
X2	162.14	252.92	186.69	-0.30	-0.33	1.56	2
X3	8.42	242.90	18.41	0.03	-0.35	28.86	1
X4	-249.72	235.78	-259.97	-0.08	-0.28	-0.94	0
X5	-31.90	250.41	-38.93	0.06	-0.38	-7.85	1

X6	-686.41	1044.48	-733.86	0.22	-0.40	-1.52	0
X7A	0.29	0.46	0.00	0.91	-1.17	1.56	0
X7B	0.36	0.48	0.00	0.57	-1.69	1.33	0
X7C	0.35	0.48	0.00	0.64	-1.59	1.38	0

Statystyki zbioru treningowego i testowego w znacznej większości są do siebie zbliżone, co świadczy o właściwym podziale zbioru bazowego. Kolejnym krokiem jest zbadanie korelacji pomiędzy zmiennymi w obu zbiorach.



Korelacje między zmiennymi są niemal identyczne w obu podzbiorach. Możemy zauważyć niewielkie różnice w niektórych wartościach, jednak są one bardzo małe, więc w obu zbiorach występują zbliżone zależności, co jest kluczowe w tworzeniu i testowaniu modelu ekonometrycznego.

2.3. Transformacje na zbiorze danych.

Zaprezentowany zbiór danych wymaga kilku transformacji, które umożliwią przeprowadzenie badania.

Najważniejszą z nich jest przekształcenie zmiennej kategorycznej X7 na zmienne binarne (zero-jedynkowe). Nowe zmienne przyjmują wartość 1 w obserwacji, dla której wartość pierwotnej zmiennej kategorycznej odpowiadała ich kategorii. Zbieg ten jest konieczny, gdyż pracując na zmiennej jakościowej nie uwzględniona zostały by różnice pomiędzy jej wartościami. Spowodowało by to niewłaściwą estymację modelu i w rezultacie fałszywe wyniki.

Inną transformacją zastosowaną dla całego zbioru danych jest ich standaryzacja. Jej potrzeba wynika ze dużych różnic jakie przyjmują wartości poszczególnych cech.

Standaryzacja polega na odjęciu od wartości średniej wszystkich wartości zmiennej oraz podzieleniu przez jej odchylenie standardowe. W skutek tego otrzymywana jest zmienna o średniej wartości oczekiwanej zero i odchyleniu standardowym 1.

Kolejnym krokiem jest poradzenie sobie z brakami danych. W związku z faktem, że zbiór danych z którego korzystano w tym projekcie ma dużo obserwacji, skorzystano z metody usunięcia każdej obserwacji, która zawierała jakiekolwiek brak w danych. Ostatecznie usunięto 24 wiersze.

Jako że wykresy pudełkowe przedstawiły znaczną liczbę wartości odstających, zwłaszcza dla zmiennych objaśniających, konieczne jest pozbycie się takich obserwacji, gdyż mogą zaburzać estymację parametrów modelu. Ta operacja spowodowała zmniejszenie się zbioru danych o 53 elementy.

2.4. Dobór zmiennych do modelu i wybór ostatecznej postaci analitycznej

Początkowo utworzony model zawiera wszystkie zmienne, które pojawiły się w zbiorze danych. Wykonany test Shapiro-Wilka pozwala stwierdzić normalność reszt modelu, dzięki czemu można interpretować wartości testu t-studenta. Wynika z nich, że zmienna X3 jest nieistotna na poziomie istotności 5%. Wartość parametru dla zmiennej X6 jest niemożliwa do wyestymowania ze względu na skrajnie wysoką korelację tej zmiennej ze zmienną X2 (współliniowość). Potwierdzeniem tego są przedstawione we wcześniejszym podrozdziale tablice korelacji. Niezbędne w tym przypadku będzie pozbycie się tej zmiennej.

	Parametr	Błąd	Wartosc t	P(>|t|)	
stała	0.81697	0.03135	26.061	< 2e-16	***
X1	0.11585	0.01788	6.481	1.74e-10	***
X2	-0.08100	0.01574	-5.146	3.48e-07	***
X3	-0.01812	0.01456	-1.245	0.2136	
X4	0.19349	0.01843	10.498	< 2e-16	***
X5	0.03058	0.01462	2.091	0.0369	*
X6	NA	NA	NA	NA	
X7A	-1.68680	0.05207	-32.395	< 2e-16	***
X7B	-0.83477	0.04079	-20.466	< 2e-16	***
Współczynnik detrminacji: 0.8614					
Test Shapiro-Wilka: p-value = 0.7455					
Test reset p-value = 1.173e-07					

Pomimo usunięcia zmiennej X6, model nie spełnia założenia o liniowości postaci modelu. Dowodem na to jest test RESET, którego p-value na poziomie 1.173e-07 jednoznacznie wskazuje nie odrzucenie hipotezy zerowej. W tym przypadku należy podjąć próbę modyfikacji zmiennych w modelu korzystając z wybranych transformacji m.in. logarytmowania, potęgowania, wymnażania wybranych zmiennych itp. Mimo podjętych licznych prób przekształceń postaci modelu, uzyskane wyniki nie były zadowalające (brak poprawy postaci modelu, spadek istotności parametrów). Poniżej przedstawiono wybrane z nich:

	Parametr	Błąd	Wartosc t	P(>|t|)	
stała	1,011	0,035	28,625	< 2e-16	***
X1^2	0,015	0,013	1,162	0.245	
X2	-0,088	0,018	-5,030	6.27e-07	***
X3	-0,022	0,016	-1,347	0.178	
X4^2	0,011	0,012	0,882	0.378	
X5	0,035	0,016	2,124	0.034	*
X7A	-2,135	0,044	-49,061	< 2e-16	***
X7B	-1,061	0,041	-25,584	< 2e-16	***
Współczynnik detrminacji: 0.826					
Test Shapiro-Wilka: p-value = 0.02					
Test reset p-value 0.017					

	Parametr	Błąd	Wartosc t	P(>|t|)	
stała	0,353	0,026	13,781	< 2e-16	***
X1*X7B	0,057	0,041	1,385	0.167	
X2	-0,151	0,021	-7,195	1.64e-12	***
X3	-0,047	0,020	-2,374	0.018	*
X4	0,362	0,023	15,606	< 2e-16	***
X5	0,071	0,020	3,611	0.000327	***
X7A	-1,125	0,051	-22,183	< 2e-16	***
Współczynnik detrminacji: 0.7435					
Test Shapiro-Wilka: p-value = 0,002					
Test reset p-value 6.294e-07					

Ostatecznym rozwiązaniem w tym przypadku jest pozbycie się zmiennych binarnych, utworzonych na podstawie zmiennej kategorycznej. W wyniku tego zabiegu zmienne, które mogą zostać użyte w modelu to X1, X2, X3, X4, X5. Poniżej znajduje się model wykorzystujące wszystkie te zmienne.

	Parametr	Błąd	Wartosc t	P(>|t|)	
stała	-0.00741	0.02280	-0.325	0.74532	
X1	0.34441	0.02608	13.206	< 2e-16	***
X2	-0.22024	0.02408	-9.146	< 2e-16	***
X3	-0.04776	0.02303	-2.074	0.03843	*
X4	0.48101	0.02569	18.724	< 2e-16	***
X5	0.06668	0.02316	2.879	0.00412	**
Współczynnik detrminacji: 0.6487					
Test Shapiro-Wilka: p-value = 0.4748					
Test reset p-value = 0.5165					

Według wyników estymacji wszystkie zmienne modelu są istotne na wysokim poziomie. Dodatkowo nie ma problemów z normalnością reszt oraz liniowością postaci modelu. Jednakże do ostatecznego wyboru czy wszystkie ze zmiennych powinny znaleźć się w modelu wykorzystano metodę Hellwiga. Dzięki niej dla wszystkich możliwych kombinacji oszacowana została pojemność informacyjna. W tabeli poniżej przedstawiono 10 najwyższych wyników.

X1	X2	X3	X4	X5	h
+	+		+	+	0.741
+			+	+	0.732
+	+	+	+	+	0.723
			+	+	0.710
+		+	+	+	0.709
	+		+	+	0.707
	+	+	+	+	0.678
		+	+	+	0.670
+	+			+	0.661
+				+	0.651

Według metody Hellwiga najlepszą opcją jest wybranie zmiennych X1, X2, X4 i X5. Jednak dołożenie zmiennej X3 nie zmniejsza pojemności wyraźnie, dlatego warto jest zachować i tą zmienną. Dodatkowo zmienna, jak zostało to wcześniej przedstawione jest istotna statystycznie, a przy doborze zmiennych warto zawsze kierować się chęcią pozostawienia jak największej ich liczby. Postępowania takie może przynieść korzyści w dalszej części, w postaci bardziej rozwiniętej interpretacji otrzymanych wyników.

2.5. Diagnostyka poprawności doboru modelu

Tworząc model ekonometryczny korzystamy z metody najmniejszych kwadratów do szacowania parametrów modelu. Aby model był poprawnie stworzony oraz aby móc poprawnie weryfikować otrzymane wyniki konieczne jest sprawdzenie założeń MNK. W celu dokonania prognozy na podstawie stworzonego modelu należy również sprawdzić pewne założenia. Poniżej znajdują się wszystkie testy, które diagnozują poprawność doboru modelu.

Składnik losowy ma wielowymiarowy rozkład normalny

Jednym z założeń metody najmniejszych kwadratów jest normalność rozkładu składnika losowego. Założenie to jest istotne przy wnioskowaniu statystycznym. Szacując parametry modelu MNK możemy szacować ich istotność za pomocą testu t studenta. Wyniki te są wiarygodne tylko w przypadku gdy składnik losowy ma rozkład normalny, a więc spełnienie tego założenia jest kluczowe w celach poprawnego odczytania uzyskanych wyników.

Do badania normalności rozkładu istnieje wiele różnorodnych testów skupiających się na różnych aspektach i mających swoje wady i zalety. Przykładowe testy to: test Andersona-Darlinga, test Kołmogorowa-Smirnowa, test Lillieforsa czy test Pearsona. W celu zbadania normalności rozkładu reszt rozpatrywanego modelu skorzystano z testu Shapiro-Wilka.

Shapiro-WIlka test:
H0: Próba pochodzi z populacji o rozkładzie normalnym
H1: Próba nie pochodzi z populacji o rozkładzie normalnym.

Wartość p-value testu Shapiro-Wilka przeprowadzonego na składniku losowym badanego modelu wynosi 0,47. Jest to wartość wyższa niż przyjęty poziom istotności alfa (0,05). Nie ma podstaw do odrzucenia hipotezy zerowej – próba pochodzi z populacji o rozkładzie normalnym. Założenie o normalności rozkładu składnika losowego zostało spełnione.

Rząd macierzy X równy jest liczbie szacowanych parametrów

Rząd macierzy jest to liczba liniowo niezależnych kolumn (kolumny odzwierciedlają zmienne objaśniające zbioru danych). Założenie to zapewnia, że estymatory można wyznaczyć w sposób jednoznaczny i poprawny.

Założenie odnośnie rządu macierzy jest ściśle połączone z dwoma kolejnymi założeniami. Założenie to sprowadza się do zbadania współliniowości. Współliniowość zbadano w dalszej części pracy.

Liczebność próby jest większa niż liczba szacowanych parametrów

Liczba obserwacji musi być co najmniej równa liczbie szacowanych parametrów. Założenie to jest ważne gdyż pozwala na otrzymanie poprawnych wyników w modelu. Gdy liczba obserwacji jest zbyt mała może prowadzić to do przeszacowania modelu oraz zbytniego dopasowania modelu do danych. Spowoduje to, że model będzie się sprawdzał tylko dla danych na których był tworzony, a wyestymowane zależności nie będzie można generalizować i przenosić na inny zbiór danych.

W rozpatrywanym modelu istnieje 5 zmiennych oraz kilkaset obserwacji. Tak duży zbiór obserwacji pozwoli na stworzenie dobrego modelu. Założenia MNK o liczebności próby jest spełnione.

Nie występuje zjawisko współliniowości pomiędzy zmiennymi objaśniającymi

Współliniowość oznacza zbyt mocne skorelowanie zmiennych wprowadzonych do modelu. Zmienne mocno skorelowane przenoszą podobne informacje – wprowadzając takie zmienne do modelu może spowodować złe oszacowanie parametrów.

Jedną z metod szacowania współliniowości jest parametr VIF. Parametr ten obliczany jest dla każdej zmiennej objaśniającej w modelu, dzięki temu możliwa jest identyfikacja zmiennej, która powoduje współliniowość.

W przypadku wykrycia współliniowości należy usunąć z modelu zmienną, która jest przyczyną współliniowości (a więc jedną ze zmiennych, które są ze sobą wysoko skorelowane). Zmienne te przenoszą tą samą informację. Można również skorzystać z analizy grzbietowej, sztucznie zmniejszając współczynnik korelacji aż do osiągnięcia zadawalających wyników.

W celu zbadania współliniowości w rozpatrywanym modelu skorzystano ze współczynnika wariancji inflacji VIF. VIF równe 1 oznacza brak współliniowości zmiennych. W przypadku gdy VIF jest poniżej 4 również możemy założyć brak współliniowości, natomiast dla większych wartości możemy stwierdzić, że współliniowość występuje.

Poniżej znajdują się wartości VIF dla pięciu zmiennych rozpatrywanych w modelu. Każda z wartości jest bliska 1, a więc możemy uznać, że współliniowość nie występuje.

X1	X2	X3	X4	X5
1.26	1.11	1.01	1.29	1.02

Założenie MNK o braku współliniowości zmiennych zostało spełnione.

Wartość oczekiwana składnika losowego jest równa zero

Kolejnym z założeń MNK jest zerowa wartość oczekiwana składnika losowego modelu. Założenie to jest istotne i oznacza, że zakłócenia reprezentowane przez składniki losowe wzajemnie się redukują. Założenie to jest konieczne w celu osiągnięcia estymatorów nieobciążonych.

W przypadku rozpatrywanego modelu wartość średniej składnika losowego wynosi -1.8e-17. Jest to wartość na tyle mała, że możemy uznać, że jest to zero, a więc założenie o zerowej wartości oczekiwanej składnika losowego jest spełnione.

Składnik losowy ma stałą skończoną wariancję

Stałość oraz skończoność wariancji możemy nazwać homoskedastycznością. MNK zakłada homoskedastyczność składnika losowego. Założenie to stosowane jest w celu uproszczenia modelu oraz obliczeń. W metodzie MNK założenie to potwierdza, że wartość wariancji zakłóceń nie zależy od numeru obserwacji.

W celu zbadania czy występuje homoskedastyczność czy heteroskedastyczność możemy skorzystać z kilku testów np. test Harrisona-Mcabe'a czy test White'a. W przeprowadzonym badaniu skorzystano z testu Breuscha-Pagana. Jest to jeden z najpopularniejszych testów, szczególnie przydatny w przypadku, gdy model ma kilka zmiennych.

Breusch-Pagan (BP) test:
H0: występuje homoskedastyczność
H1: występuje heteroskedastyczność

Wartość p-value testu BP przeprowadzonego na składniku losowym rozpatrywanego modelu wynosi 0,93. Wartość ta jest większa niż przyjęty poziom istotności alfa (0,05), a więc nie ma podstaw do odrzucenia hipotezy zerowej. Rozpatrywany model ma składnik losowy o stałej oraz skończonej wariancji. Założenie MNK o występowaniu homoskedastyczności zostało spełnione.

Nie występuje zjawisko autokorelacji składnika losowego

Autokorelacja jest skorelowanie zmiennej z tą samą zmienna z innego obiektu/okresu. W przypadku występowania zjawiska autokorelacji składnika losowego macierz wariancji-kowariancji nie jest macierzą diagonalną, a więc nie możemy skorzystać z metody MNK. W takiej sytuacji można zastosować uogólniona metodę MNK (UMNK).

W przypadku wystąpienia autokorelacji składnika losowego estymatory MNK przestają być najefektywniejsze Autokorelacja składnika losowego wiąże się z niewłaściwą postacią modelu lub pominięciem w modelu istotnej zmiennej.

Jednym z popularniejszych testów na zbadanie autokorelacji jest test Durbina-Watsona. Test ten estymuje współczynnik autokorelacji i w zależności od wartości statystyki d możemy przyjąć lub odrzucić hipotezę zerową.

Durbin-Watson (DW) test:
H0: brak autokorelacji
H1: występuje autokorelacja I rzędu

Wartość p-value testu DW dla rozpatrywanego modelu wynosi 0.93, a więc jest większa od poziomu istotności alfa (0,05). Nie ma podstaw to odrzucenia hipotezy zerowej, a więc współczynnik autokorelacji jest nieistotny. Założenie MNK o braku autokorelacji składnika losowego jest spełnione.

Szacowany model ekonometryczny jest liniowy względem parametrów

Kolejnym ważnym założeniem MNK jest liniowość modelu względem parametrów. Założenie to jest kluczowe, gdyż estymatory metody MNK są liniowe. W celu sprawdzenia tego założenia skorzystano z testu RESET Ramseya.

Test RESET sprawdza poprawność specyfikacji dla modelów regresji liniowej. Sprawdza on czy liniowa postać modelu względem funkcji kwadratowej czy szeciennej jest najlepsza. Test ten nie porównuje wybranego modelu do alternatywnych wersji, a więc nie sugeruje on lepszej postaci modelu.

Test RESET:
H0: liniowa postać modelu jest właściwa
H1: liniowa postać modelu nie jest właściwa

Wartość p-value testu RESET wynosi 0,15 i jest większa niż przyjęty poziom istotności. Nie ma podstaw do odrzucenia hipotezy zerowej. Założenie MNK o liniowości modelu względem parametrów jest spełnione.

Musi występować odpowiednio wysoki stopień dopasowania modelu (dopasowanie modelu do danych empirycznych)

Badanie poziomu dopasowania modelu do danych empirycznych ma na celu weryfikacje, czy model ten w wystarczającym stopniu wyjaśnia kształtowanie się zmiennej objaśnianej. Miary służące do weryfikacji tego założenia to: odchylenia standardowe reszt, współczynnik zmienności losowej czy kryteria informacyjne AIC, BIC.

W celu zbadania dopasowania modelu wykorzystano współczynnik determinacji – informuje on o tym, jaka część zmienności zmiennej objaśnianej w próbie pokrywa się z korelacjami zmiennych zawartych w modelu. Współczynnik determinacji przyjmuje wartości z przedziału od 0 do 1. W celu badania istotności współczynnika determinacji skorzystano ze statystyki F.

Wartość współczynnika determinacji dla rozpatrywanego modelu wynosi 0,65, a p-value testu istotności współczynnika determinacji wynosi 2.2e-16. Oznacza to, że wartość współczynnik jest istotna. Założenie MNK o wysokim syopniu dopasowania modelu jest spełnione.

Stabilność modelu

Sprawdzenie stabilności oszacowań parametrów modelu ekonometrycznego jest konieczne w przypadku analizy strukturalnej jak i w celu przeprowadzenia prognozowania. Do zbadania stabilności modelu możemy skorzystać z testu QLR czy testu Chowa.

Test Chowa służy do sprawdzenia czy parametry modelu będą takie same dla kilku różnych podpróbek. Brak stabilności modelu podważa model – interpretacja oszacowanych parametrów może być błędna. Brak stabilności wpływa również na niemożność udowodnienia, że estymatory są nieobciążone czy efektywne

Test Chowa:
H0: stabilność parametrów
H1: model nie jest stabilny

P-value testu Chowa dla badanego modelu wynosi 0,38 – jest większe niż przyjęty poziom istotności (0,05), a więc nie ma podstaw do odrzucenia hipotezy zerowej. Założenie o stabilności modelu zostało spełnione. Możliwe jest przeprowadzenie prognozowania przy użyciu rozpatrywanego modelu.

Koincydencja

Model jest koincydentny jeżeli każda zmienna w modelu jest koincydentna. W celu sprawdzenia koincydencji należy zbadać współczynnik korelacji zmiennej objaśnianej z daną zmienną objaśniającą oraz współczynnik modelu ekonometrycznego przy danej zmiennej. Jeżeli znaki współczynników są sobie równe model jest koincydentny.

W przypadku gdy model nie jest koincydenty należy dobrać inne zmienne do modelu. Koincydencja ocenia sensowność wyestymowanych parametrów, a model niekoincydentny

nie powinien być wykorzystywany. Warto zwrócić uwagę na współliniowość modelu, która może być przyczyną braku koincydencji.

	współczynnik	korelacja ze zm. Y	koincydencja
X1	0.60	0.34	TRUE
X2	-0.44	-0.22	TRUE
X3	-0.13	-0.05	TRUE
X4	0.69	0.48	TRUE
X5	0.17	0.07	TRUE

Analizując wszystkie pięć zmiennych, które wykorzystano w modelu możemy stwierdzić, że model jest koincydentny. Każda ze zmiennych ma taki sam znak oszacowanego parametru jak i współczynnika korelacji ze zmienną objaśnianą. Założenie o koincydencji modelu jest spełnione.

Losowość próbki

Tworząc model ekonometryczny ważne jest dla nas możliwości jego generalizacji, a więc dane muszą być odpowiednio dobrane. Aby próbka była reprezentatywna dla całej populacji musi być losowa.

W celu zbadania losowości zmiennych skorzystano z testu Walda-Wolfitza (testu serii). Jest to nieparametryczny test, który sprawdza czy każdy element w zbiorze jest niezależną próbką z rozkładu.

Test serii:
H0: dobór jednostek do próby jest losowy
H1: dobór jednostek do próby jest nielosowy

	Y	X1	X2	X3	X4	X5
p-value	0.94	0.22	0.25	0.54	1.00	0.59

P-value dla wszystkich zmiennych jest powyżej ustalonego poziomu istotności. Nie mamy podstaw do odrzucenia hipotezy zerowej. Założenie MNK o losowości próbki jest spełnione.

Łączna istotność zmiennych

W celu zbadania łącznej istotności parametrów z modelu wykorzystuje się test Walda. Pozwala nam to stwierdzić czy w danym modelu powinniśmy pozostawić dane zmienne czy je usunąć. Test ten ma statystykę F.

Test Walda:
H0: parametry modelu są łącznie nieistotne
H1: parametry modelu są łącznie istotne

Wartość p-value testu Walda wynosi 2.2e-16. Wartość ta jest niska i pozwala nam na odrzucenie hipotezy zerowej. Parametry modelu są łącznie istotne, a więc zmienne zostały prawidłowo dobrane do modelu.

Podsumowanie

Wyniki wszystkich przeprowadzonych testów są pozytywne i pozwalają stwierdzić poprawność doboru modelu. Założenia MNK są spełnione, a wszystkie założenia, które dodatkowo sprawdzono jednoznacznie stwierdzają, że model jest właściwy.

2.6. Interpretacja parametrów modelu.

Ostateczna postać modelu wraz z oszacowanymi wartościami parametrów prezentuje się następująco:

$$Y = -0.007 + 0.344X1 - 0.220X2 - 0.048X3 + 0.481X4 + 0.067X5$$

Wyraz wolny modelu jest bliski wartości zero. Nie ma on istotnego wpływu na zmienną objaśnianą. Największa wartość współczynnika znajduje się przy zmiennej X4 i wynosi aż 0,48. Warto pamiętać, że dokonano skalowania wartości zmiennych, a więc wzrost X4 o jednostkę nie powoduje wzrostu Y o 0.48, jednak można stwierdzić, że spowoduje znaczny wzrost Y.

Kolejne wysokie współczynniki znajdują się przy zmiennych X1 oraz X2. Mają one znaczny wpływ na zmienną objaśnianą. Wzrost zmiennej X1 powoduje wzrost Y, a wzrost zmiennej X2 powoduje zmniejszenie wartości Y.

Zmienne, które mają najmniejszy wpływ na zmienną objaśnianą to zmienne: X3 oraz X5. Wzrost zmiennej X3 powoduje niewielkie zmniejszenie wartości Y, a wzrost zmiennej X5 powoduje zwiększenie się zmiennej objaśnianej.

Podsumowując, zmienne X2 oraz X3 mają negatywny wpływ na wartość zmiennej Y, a zmienne X1, X4, X5 oraz stała pozytywny wpływ.

2.7. Prognoza EX POST

Na podstawie otrzymanego modelu ekonometrycznego można przeprowadzić prognozę EX POST, gdyż jego diagnostyka nie wykazała żadnych nieprawidłowości. Szczególnie ważne w przypadku predykcji jest, że model spełnia założenie stabilności parametrów rozkładu składnika losowego (test Chowa) oraz ma jednoznacznie określoną postać liniową. Współczynnik determinacji jest na przyzwoitym poziomie około 70%, co również jest konieczne dla uzyskania korzystnych wyników oszacowań.

Wykonanie prognozy polega na podstawieniu – w miejsce wyróżnionych zmiennych objaśniających – odpowiednich danych liczbowych i wykonaniu określonych działań algebraicznych, zgodnie z analityczną postacią modelu:

$$Y = -0.007 + 0.344X1 - 0.220X2 - 0.048X3 + 0.481X4 + 0.067X5$$

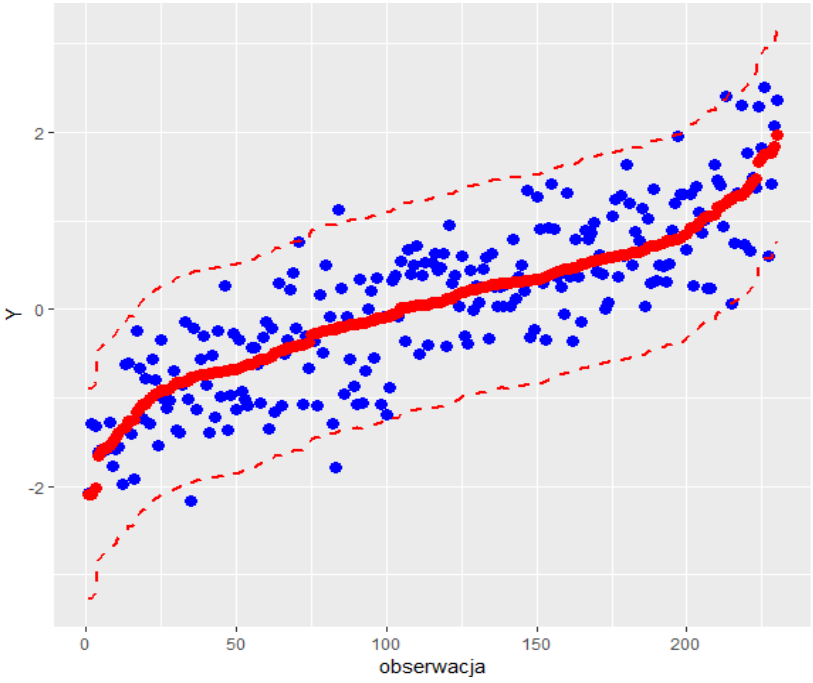
W wyniku zastosowania wzoru dla danych testowych uzyskano następujące wartości oszacowań zmiennej objaśnianej Y oraz obliczono bezwzględną różnice pomiędzy nimi, a wartościami rzeczywistymi. Część rezultatów zaprezentowano w poniższej tabeli:

X1	X2	X3	X4	X5	Y	Y prognoza	bezwzględna różnica
-0,385	0,423	-1,620	-1,209	0,526	-0,990	-0,702	0,288
-1,780	1,905	-1,088	1,713	0,080	0,334	-0,159	0,493
2,103	-2,234	1,184	1,362	-1,502	1,831	1,707	0,124
-0,681	1,335	-1,281	-1,803	-1,728	-1,589	-1,458	0,131
0,016	0,118	0,682	-0,525	0,879	0,165	-0,254	0,419
-1,558	0,284	0,521	-0,913	0,813	-1,295	-1,017	0,278
0,244	0,030	-0,342	-1,721	0,939	-0,265	-0,679	0,414
-1,038	1,469	1,119	-0,949	-1,053	-0,602	-1,268	0,666
-2,409	-0,330	0,304	0,098	0,254	-1,219	-0,715	0,504
0,457	-0,898	0,973	0,769	0,647	0,296	0,715	0,419
-0,131	0,463	-0,510	0,845	0,078	0,027	0,281	0,254
0,226	0,032	0,390	0,657	-1,809	0,080	0,240	0,16
…	…	…	…	…	…	…	…

Przedziały ufności dla prognozy wyglądają następująco:

Y	przedział lewy	przedział prawy
-0,990	-1,883	0,479
0,334	-1,349	1,032
1,831	0,519	2,895
-1,589	-2,642	-0,273
0,165	-1,432	0,924
-1,295	-2,197	0,164
-0,265	-1,860	0,503
-0,602	-2,449	-0,088
-1,219	-1,899	0,469
0,296	-0,464	1,894
0,027	-0,897	1,460
0,080	-0,940	1,420
...	...	...

Na podstawie przedstawionych wyników można stwierdzić, że model dobrze dopasowuje się do danych. Według wyznaczonego przydziału ufności 95% obserwacji mieści się w odległości około 1 jednostki od prognozy.



Więcej informacji można uzyskać dzięki następującym miarom oceny prognozy:

Miara	Wartość
Średni błąd predykcji (ME)	0,021
Średni bezwzględny błąd predykcji (MAE)	0,434
Średniokwadratowy błąd predykcji (MSE)	0,282
Średni bezwzględny błąd procentowy predykcji (MAPE)	67,142

Średni błąd predykcji w tym przypadku jest mało przydatny, gdyż zmienne przyjmują wartości dodatnie i ujemne, co zaburza obliczenia średniej z błędów.

Dużo więcej informacji pozwala uzyskać średni bezwzględny błąd predykcji. Rzeczywista wartość zmiennej Y średnio różni się od prognozy o niecałe 0,5 jednostki. Różnica ta wydaje się być wysoka, biorąc pod uwagę, że Y waha się w granicach od -2,5 do 2,5.

Błąd średniokwadratowy dobrze interpretowalny jest w zestawieniu z odchyleniem standardowym reszt modelu. W tym przypadku wynosi ok. 0,28 i jest mniejszy od odchylenia reszt (0,53). To oznacza, że wektor prognoz można uznać za zadowalający.

MAPE wynosi około 67%. Wyraża on średnią wielkość błędów prognoz dla okresu testowego w procentach. Wykorzystuje się go do porównywania różnych modeli

3. Wnioski

Stworzenie modelu ekonometrycznego pozwoliło na zgłębienie wielu aspektów, które towarzyszą tej procedurze. Zaczynając od dokładnej analizy zgromadzonych danych, przedstawieniu dokładnych statystyk oraz wykresów takich jak wykresy pudełkowe, histogramy czy wykresy zależności możliwe było szczegółowe zapoznanie się z obserwacjami. Już na takiej podstawie można było zauważyć pewne zależności, które później pozwoliły na dobór odpowiednich zmiennych do modelu jak na przykład wysoka korelacja zmiennych ze zmienną objaśnianą czy wysoka korelacja dwóch zmiennych objaśniających, która jednoznacznie wskazywała liniowość. Analiza statystyk pozwoliła zauważyć jakie transformacje potrzebne są w modelu jak na przykład usunięcie wartości odstających, usunięcie braków danych czy zamienienie zmiennej kategorycznej na zmienne jedynkowej.

Kolejnym krokiem było tworzenie modelu ekonometrycznego oraz dobór zmiennych w taki sposób, żeby miały one istotny wpływ na model oraz żeby ostateczny model miał postać liniową. Potrzebne było sprawdzenie różnych kombinacji, aby móc wybrać najlepszą postać modelu. Tworząc model należy sprawdzić wszystkie założenia stosowanych metod. Jest to o tyle istotne, że brak weryfikacji może wpłynąć na błędną interpretację nieprawdziwych wartości. Odpowiednia diagnostyka modelu pozwala na interpretację oszacowań. Na podstawie wyestymowanych parametrów modelu można określić nie tylko charakter zależności pomiędzy zmiennymi egzogenicznymi, a zmienna endogeniczna, ale również podać szczegółowo w jakim stopniu każda z nich wpływa na Y.

Dobrze skonstruowany, stabilny model pozwolił na dokonanie predykcji na zbiorze testowym. Porównanie wartości prognozowanych oraz rzeczywistych, także błędy oceny prognozy pokazały, że model umożliwia właściwe przewidywanie wartości zmiennej objaśnianej. Wynika to z jego odpowiedniego dopasowania do danych. Badanie jest potwierdzeniem tego, że przy prognozowaniu ważne jest, by model spełniał wszystkie niezbędne założenia.

W wykonanym projekcie widoczne jest, że konstrukcja modelu ekonometrycznego jest bardzo skomplikowanym procesem, wymagającym wykonania wielu dodatkowych zadań. W jego ramach przeprowadzana jest szczegółowa analiza, w wyniku której istotnie modyfikowany jest przedmiot badania. Jako rezultat całej procedury budowy modelu ekonometrycznego, otrzymywana jest solidna porcja informacji, które niosą za sobą, kluczowe z punktu widzenia tematu badania, wnioski.